Title

The Molecular Mechanisms of Transcriptional Regulatory Network Divergence

Author

Baker, Christopher Robert

Publication Date

2012

The Molecular Mechanisms of Transcriptional Regulatory Network Divergence

by

Christopher R. Baker

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Genetics

in the GRADUATE

DIVISION of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

Acknowledgements

No graduate project is truly the work of just a single person. In my work, I benefitted from collaborations with my peers, mentorship from some of the best in the field, and friendships that made the challenging days not all that bad. I am honored to acknowledge the guidance of my graduate mentor Dr. Alexander Johnson and Thesis Committee members Dr. Hao Li and Dr. David Morgan. In the lab, I benefitted from the advice of Dr. Brian Tuch, Dr. Oliver Homann, Dr. Aaron Hernday, and Dr. Quinn Mitrovich. I am proud to have been a part of collaborations with Dr. Lauren Booth, Trevor Sorrells, Dr. Brian Tuch, and Dr. Victor Hanson Smith. I owe much to the love and support of my family, my girlfriend Linet Mera, and my wonderful classmates. Plus, it was really fun to be in an all scientist rock band during my graduate years.

The Molecular Mechanisms of Transcriptional Regulatory Network Divergence

By Christopher R. Baker

Abstract

Transcriptional regulatory networks take part in the regulation of essentially every aspect of an organism's biology. Yet, the structure of these networks can be baroque and attempts to rationalize these structures based on network engineering principles have met with mixed success. In my graduate work, I have examined how the structure of modern transcriptional regulatory networks is rooted in their evolutionary history. In this light, an understanding of transcriptional regulatory network structure based exclusively on engineering design principles may not be useful. Instead, to rationalize the structure of transcriptional regulatory networks requires the careful examination of how these structures are built using pieces of existing networks, how networks are rewired by the constant force of degenerative mutation, and how constraints on transcriptional regulatory network evolution shape the path towards the gain of new functions.

Table of Contents

1. **Introduction**

   Page 1

2. **Extensive divergence in the DNA-binding specificity of a conserved transcription regulator**

   Page 10

3. **Protein modularity, cooperative binding, and hybrid regulatory states underlie transcriptional regulatory network diversifcation**

   Page 40

4. **Increased regulatory complexity as a consequence of functional interference between gene duplicates**

   Page 120

Figure Legend

Chapter 2

Figure 1 – Page 28
Significant divergence of the **α**-specific gene *cis*-regulatory sequence between *C. albicans* and *S. cerevisiae*

Figure 2—Page 29
C. albicans Mat**α**1 activates transcription by binding to the *C. albicans* **α**-specific gene *cis*-regulatory sequences

Figure 3—Page 31
Extensive DNA-binding specificity divergence of the Mat**α**1 protein

Figure 4—Page 32
The DNA-binding specificity of the *C. albicans* Mat**α**1 protein evolved after the divergence of *S. cerevisiae* and *C. albicans*

Figure 5—Page 34
Mat**α**1 DNA-binding specificity has continued to diverge within the CTG-clade

Supplemental Figure 1—Page 35
The Mcm1 binding site within *C. albicans* **α**sg cis-regulatory sequence is necessary for Ca-Mat**α**1-dependent transcriptional activation

Supplemental Figure 2—Page 36
Species-specific binding to a second set of **α**sg cis-regulatory sequences

Supplemental Figure 3—Page 37
Functional conservation of Mat**α**1 DNA- binding specificity between S. *cerevisiae* and *K. lactis*

Supplemental Figure 4—Page 38
Evolution of the Mat**α**1 HMG DNA-binding domain

Supplemental Figure 5—Page 39
Y. lipolytica **α**sg *cis*-regulatory PSSM

Chapter 3

Figure 1—Page 73
Cell-type specification in the hemiascomycetes

Figure 2—Page 75
α2 repression of the **a**sgs evolved prior to the divergence of *Saccharomyces* and *Kluyveromyces*

Figure 3—Page 77
The *cis* and *trans*-evolution underlying the gain of a new function for **α**2

Figure 4—Page 79
The contribution of non-specific protein interactions to early intermediates

Figure 5—Page 81
Regulation of the **a**sgs in *Lachancea kluyveri*

Figure 6—Page 83
Regulation of the **a**sgs in *Kluyveromyces lactis*

Figure 7—Page 85
The gain of the hybrid regulatory state facilitated diversification of **a**sg regulation
Supplemental Figure 1—Page 91
Further support that the evolution occurred of **a**sg repression occurred in the ancestor of the *Saccharomyces-Kluyveromyces* clade and required the gain of a Tup1 and Mcm1 interaction within the **α**2 protein
Supplemental Figure 2—Page 93
The evolution of **α**2-repression exploited an ancestral Mcm1 interaction surface
Supplemental Figure 3—Page 95
Over-expression of *S. cerevisiae* **α**2 failed to repress out-group **a**-specific gene *cis*-regulatory sequences
Supplemental Figure 4—Page 96
Regulation of *L. kluyveri* **a**-specific genes
Supplemental Figure 5—Page 98
The *K. lactis* and *L. kluyveri* **a**-specific genes
Supplemental Figure 6—Page 100
Regulation of *K. wickerhamii* **a**-specific genes
Supplemental Figure 7—Page 102
The mutational path of the loss of **α**2 repression in the *Kluyveromyces*

Chapter 4

Figure 1—Page 144
The duplication and divergence of a multifunctional transcription regulator
Figure 2—Page 146
The pre-duplication ancestor complemented both paralogs in *S. cerevisiae*
Figure 3—Page 147
Duplicates were maintained through the reciprocal loss of ancestral interactions
Figure 4—Page 148
Reduced DNA-binding affinity of a post-duplication paralog

Table Legend

No tables have been included in this dissertation.

Chapter 1:

Introduction

Abstract:

Transcriptional regulatory networks take part in the regulation of essentially every aspect of an organism's biology. Yet, the structure of these networks can be baroque and attempts to rationalize these structures based on network engineering principles have met with mixed success. In my graduate work, I have examined how the structure of modern transcriptional regulatory networks is rooted in their evolutionary history. In this light, an understanding of transcriptional regulatory network structure based exclusively on engineering design principles may not be useful. Instead, to rationalize the structure of transcriptional regulatory networks requires the careful examination of how these structures are built using pieces of existing networks, how networks are rewired by the constant force of degenerative mutation, and how constraints on transcriptional regulatory network evolution shape the path towards the gain of new functions.

Transcriptional regulatory networks integrate information & make decisions

Cells must adapt to the environmental, physiological, and developmental cues they receive over their lifespan. In most cases, altering genomic DNA makes an inefficient response to these inputs. This type of change is slow, irreversible, untunable, and if the genomic alteration occurs in the germ line, it will be passed on to the progeny as well. Instead, cells have evolved a complex regulatory apparatus that modifies the activity of functional molecules in response to informational cues. These types of regulatory events possess reversibility, tunability, speed, and can impact the biology of progeny cells (epigenetics) but can also be erased away with each new generation.

A universal component of this regulatory apparatus are transcriptional regulatory networks. These networks consist of two fundamental pieces: *cis*-regulatory DNA binding sequences and sequence-specific DNA-binding proteins known as transcription regulators. In general, *cis*-regulatory sequences are short pieces of DNA (less than 50 nucleotides) that occur outside of the DNA sequences that code for functional molecules (protein and RNA). The composition and order of nucleotides within these sequences are non-random, because where a transcription regulator will bind depends on the sequence of the *cis*-regulatory site.

Transcription regulators make contacts to DNA over a short stretch of DNA (generally less than 10 nucleotides). These proteins will remain bound to some DNA sequences for far greater durations than other sequences. Additionally, these preferred binding sequences are related (separated by several nucleotide differences, but sharing many nucleotides in common). Thus, we describe transcription regulators as possessing DNA-binding specificity (they prefer to bind a certain sequence and its close relatives) and we can describe these specificities as motifs (see Figure 1, Chapter 2). Different transcription regulators will have different DNA-binding specificities, although this not always the case (Chapter 3, 4).

Once stably bound to DNA, transcription regulators recruit additional transcriptional machinery that translates this DNA binding event to a change in the transcription of a target gene. Typically, *cis*-regulatory sequences are located proximal to their target gene, producing either a repressive or activating effect on the level of transcript produced from that target gene.

In total, a transcriptional regulatory network detects a piece of input information, initiates a response (a change in the pattern of DNA-binding by transcription regulators), and then reversibly adjusts the levels of functional molecular in the cell (through the recruitment of the general transcriptional control machinery). Scale these principles up to the entire genome, which even a single cell eukaryote contains hundreds of transcription regulators and thousands of *cis*-regulatory sequences, and the vast potential of the cell's regulatory architecture begins to emerge.

Transcription regulators also tend to bind *cis*-regulatory sequences in combination with other transcription regulators, forming both protein-DNA and protein-protein interactions to other regulatory proteins bound to adjacent DNA sequences. Thus, *cis*-regulatory sequences are often collections of several binding sites for several different transcription regulators, which are in turn capable of stabilizing one-another binding through forming protein-protein interactions. For this reason, the total interaction energy for a given transcription regulator to occupy a *cis*-regulatory sequence is the sum of DNA-protein and protein-protein interactions formed at that site.

Challenges to rationalizing the structure of transcriptional regulatory networks

In our modern world, at work and at home, we depend upon electrical circuits. These electrical circuits receive input information (a signal), pass this signal through a rationally designed set of filters, capacitors, and logic gates, and output a response that these gates dictate based upon the input signal. In these three steps (detection of a signal, passage through a decision making network, output of a regulated response), the action of electrical circuits closely resembles the action of transcriptional regulatory networks. In fact, transcriptional regulatory network architectures have been discovered in several different organisms that behave in ways that parallel the action of components of electrical circuits (noise filters, logic gates, capacitors) (Alon 2007).

These are elegant examples of the parallels between engineering principles and biology. However, there are also important differences between transcriptional regulatory networks and electrical circuits. Namely, electrical circuits are rationally designed, whereas transcriptional regulatory networks are built by evolution. In an idealized situation, natural selection might chance upon the same efficient designs for transcriptional regulatory networks that characterize electrical circuits. Yet, at least, four principles can be invoked that suggest that this will rarely be the case.

First, natural selection is arguably the weakest of the four evolutionary forces (natural selection, neutral drift, mutation, and gene flow) (Lynch and Conery 2003). If transcriptional regulatory networks emerge largely in the absence of efficient natural selection, then there is no predication that the structure of these networks should mirror the efficiency and simplicity of electrical circuits. Second, there are no blank canvases in biology. Instead, natural selection can only guide the design of transcriptional regulatory networks within the framework of the existing global regulatory network architecture of cell. Existing transcriptional regulatory networks will both constrain the evolutionary paths available to the emergence of a new network and provide the substrate from which a new network will be assembled. Thus, existing regulatory networks may block optimal network design for an emerging network and baroque network architectures can emerge as a consequence of piecing together existing pieces of transcriptional regulatory networks to create new networks. Third, the demands on transcriptional regulatory networks change over time. A transcriptional regulatory network may diverge under pressure from natural selection in response to an immediate challenge facing the organism. However, with the passage of eons, that initial challenge may dissipate, leaving behind a transcriptional regulatory network that is now exploited to serve a different purpose but was shaped by natural selection to execute an entirely different function. Finally, transcriptional regulatory networks are continuously being degraded by degenerative mutations. This means that optimal network designs can be washed away over evolutionary time by degenerative mutations that slip past natural selection and leave behind network architectures that are 'just good enough to get the job done'.

The features of modern transcriptional regulatory networks are largely consistent with this chasm between rational network design and what exists in actual biological networks. Studies of transcriptional regulators governing complex traits reveal densely interconnected networks where transcription regulators form all possible interactions with the other regulators (Nobile, Fox et al. 2012). These regulators then bind 100s to 1000s of target genes, revealing the high level of interconnectivity across the entire cell. One might argue that high connectivity within a transcription regulator could be advantageous under some conditions, but the shear scale of connectivity becomes increasingly difficult to rationalize in the context of network design principals. Next, orthologous transcription regulators from different species tend to regulate remarkably few overlapping genes (Schmidt, Wilson et al. 2010) (Borneman, Gianoulis et al. 2007). This is likely a reflection of the continuous turnover of transcriptional regulatory networks through the process of degenerative mutation and the chance acquisition of new binding sequences. Finally, over evolutionary time, alternative transcriptional circuits have replaced existing structures while maintaining regulatory logic (Tsong, Tuch et al. 2006). Thus, there are multiple regulatory network architectures capable of achieving the same function and cell appears to be transitioning over time between these different solutions. Taken together, these findings strongly suggest that modern networks do not reflect optimized network structures but are rather part of a continuous cycle of turnover in transcriptional regulatory network composition and structure.

Evolution within transcriptional regulatory networks

There has been considerable interest in the evolution of transcriptional networks given both (a) the fundamental importance of transcriptional regulatory networks to the biology of the cell and organism and (b) what is considered to be the fast rate of divergence in these networks between species (Schmidt, Wilson et al. 2010) (Borneman, Gianoulis et al. 2007). Ideas about how transcriptional regulatory networks evolve have centered around two related concepts: evolvability and minimizing pleiotropy.

The evolvability of transcriptional regulatory networks describes the capacity of just a few mutations to lead to a rewiring event within the network. For instance, the emergence of a new DNA-binding site for a transcription regulator can require only a handful of mutations (Chapter 3). Minimized pleiotropy describes how evolution within transcriptional regulatory networks can proceed through trajectories that do not compromise organismal fitness. For instance, the gain or loss of *cis*-regulatory sequences may alter the expression of a target gene under a certain condition, but it will not alter that target gene's biology under all possible conditions (Carroll 2005).

Both the examples cited above address changes in *cis*-regulatory sequence, which have been the primary interest of studies in the evolution of transcriptional regulatory networks (Carroll 2005). However, as will be addressed in Chapters 2-4, transcription regulators also experience important evolutionary transitions. The principles of evolvability and minimized pleiotropy that apply to *cis*-regulatory sequence evolution also apply to the evolution of transcription regulators. In fact, large-scale network rewiring events may depend on evolution in transcription regulators to 'jump start' these global events (Chapter 3).

In my graduate work, I have attempted to build mechanistic examples that illustrate how evolutionary history shapes the structure of modern transcriptional networks. I have worked in the hemiascomycete yeast (mostly in the model hemiascomycete yeast *S. cerevisiae*). The experimental strategy applied across much of my work has been the reconstitution of components of transcriptional regulatory networks from other hemiascomycetes in the model yeast *S. cerevisiae*. The work has also employed computational, biochemical, and genetic experiments in different yeast species. I believe the work illustrates at least five core principles:

1) The importance of neutral drift to reshaping transcriptional regulatory network structure (Chapters 2-4)
2) The exploitation of existing features of transcriptional regulatory networks to evolve new functions (Chapter 3)
3) The constraints on transcriptional regulatory network evolution have shaped the structure of modern networks (Chapters 3-4)
4) The protein modularity and cooperative interactions of transcription regulators can facilitate network rewiring events (Chapters 2-4)
5) The ancient constraints that shaped the structure of modern networks can be disguised by subsequent evolutionary events (Chapters 3-4)

In conclusion, the intention of the work is to influence the way molecular biologists understand their own studies. The properties of biological systems at the molecular level are shaped by an evolutionary history that is not always intuitive. Applying anthropomorphic constructs, such as engineering principles, to understand these molecular systems will at times lead to false conclusions. The properties of a regulatory network within a given model species should not be understood as the final product of a rational design process. Rather, it is a snapshot in evolutionary time. Even in the absence of new extrinsic pressures, that regulatory network will diverge (often dramatically) from its ancestral state. Yet, this critique is not meant to imply we are incapable of discovering the reasons for why networks are structured in the way they are.

Instead, it is essential to look beyond modern networks and into the past. Evolution does not expunge all the features of ancestral networks; instead, remnants of the history that shaped these network remains intact, providing the answers we seek.

Bibliography

Alon, U. (2007). An introduction to systems biology : design principles of biological circuits. Boca Raton, FL, Chapman & Hall/CRC.

Borneman, A. R., T. A. Gianoulis, et al. (2007). "Divergence of transcription factor binding sites across related yeast species." Science **317**(5839): 815-819.

Carroll, S. B. (2005). "Evolution at two levels: on genes and form." PLoS Biol **3**(7): e245.

Lynch, M. and J. S. Conery (2003). "The origins of genome complexity." Science **302**(5649): 1401-1404.

Nobile, C. J., E. P. Fox, et al. (2012). "A recently evolved transcriptional network controls biofilm development in Candida albicans." Cell **148**(1-2): 126-138.

Schmidt, D., M. D. Wilson, et al. (2010). "Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding." Science **328**(5981): 1036-1040.

Tsong, A. E., B. B. Tuch, et al. (2006). "Evolution of alternative transcriptional circuits with identical logic." Nature **443**(7110): 415-420.

Chapter 2:

Extensive DNA-Binding Specificity Divergence of a Conserved Transcription Regulator

Abstract:

The DNA sequence recognized by a transcription regulator can be conserved across large evolutionary distances. For example, it is known that many homologous regulators in yeasts and mammals can recognize the same (or closely related) DNA sequences. In contrast to this paradigm, we describe a case where the DNA-binding specificity of a transcription regulator has changed so extensively (and over a much smaller evolutionary distance) that its *cis*-regulatory sequence appears unrelated in different species. Bioinformatic, genetic, and biochemical approaches were used to document and analyze a major change in the DNA-binding specificity of Matα1, a regulator of cell-type specification in ascomycete fungi. Despite this change, Matα1 controls the same core set of genes in the hemiascomycetes because its DNA recognition site has coevolved with it, preserving the protein-DNA interaction but significantly changing its molecular details. Matα1 and its recognition sequence diverged most dramatically in the common ancestor of the CTG-clade (*C. albicans, C. lusitaniae,* and related species), apparently without the aid of a gene duplication event. Our findings suggest that DNA-binding specificity divergence between orthologous transcription regulators may be more prevalent than previously thought and that seemingly unrelated *cis*-regulatory sequences can nonetheless be homologous. These findings have important implications for

understanding transcriptional network evolution and for the bioinformatic analysis of regulatory circuits.

Introduction:

The importance of changes in the DNA-binding specificity of orthologous transcription regulators to the evolution of transcriptional networks is an open question. Several lines of evidence have been used to argue that divergence in transcription regulator DNA-binding specificity occurs infrequently. These arguments include the amino-acid conservation of transcription regulator DNA-binding domains *(1),* the potentially pleiotropic nature of alterations to transcription regulator DNA-binding specificity *(2)*, and the conservation of function across large evolutionary distances for certain transcription regulators *(3,4)*. Several cases of drift in the transcription regulator DNA-binding specificity have been documented across species, but in these cases the changes were limited to a small number of amino-acid positions and the *cis*-regulatory sequence remains similar across species *(5,6)*. Here we show that the DNA-binding specificity of a deeply conserved transcription regulator (Matα1) can change so extensively that its *cis*-regulatory sequence in different species appears unrelated as assessed by bioinformatic criteria.

In the model yeast *S. cerevisiae,* the HMG DNA-binding domain transcription regulator Matα1 activates a set of genes involved in cell-type (mating-type) specification, known as the α-specific genes. Matα1 associates with α–specific gene promoters through direct, sequence specific DNA-binding aided by a protein-protein interaction with a second sequence specific DNA-binding protein, Mcm1 *(7,8)*. This basic form of α–specific gene regulation appears to be conserved in the pathogenic yeast *C. albicans*, which is estimated to have diverged between 100 and 300 mya from the lineage that gave rise to *S. cerevisiae (9)*. For example, deletion of the Matα1 ortholog in *C. albicans* results in a loss of α–specific gene expression, and the *C. albicans* Mcm1 ortholog has been shown to bind α-specific gene promoters *(10,11)*. Despite the overall similarity of the regulatory scheme, the *cis*-regulatory DNA sequences that regulate the α-specific genes have diverged substantially between the two yeasts *(11)*. Here we demonstrate that the source of this divergence is the evolution of a new Matα1 DNA-binding specificity.

Results:

Significant divergence of the α-specific gene *cis*-regulatory sequence between *C. albicans* and *S. cerevisiae*

To computationally demonstrate the divergence of the α–specific gene *cis*-regulatory DNA sequences between *C. albicans* and *S. cerevisiae*, position specific scoring matrices (PSSMs) for α-specific gene *cis*-regulatory sequences were computed for the *S. cerevisiae* and *C. albicans* clades (Figure 1A). For this study we define the *S. cerevisiae* clade as encompassing *S. cerevisiae, S. bayanus, S. mikatae,* and *S. paradoxus (12)* and the *C. albicans* clade as *C. albicans, C. tropicalis,* and *C. dubliniensis (13).* The extent of divergence between the two PSSMs was then measured, revealing significant differences between the α–specific gene *cis*-regulatory sequences from the *C. albicans* and *S. cerevisiae* clades (Figure 1B). Although the Mcm1 binding site was strongly conserved between the two clades (E-value: .0016; see methods), the adjacent sequence (known to be recognized by Matα1 in *S. cerevisiae*) was not conserved (E-value: >1200). Instead, the *C. albicans* clade appeared to have a different binding site in the same position.

At least three models can be invoked to explain this divergence. In the first model, “regulatory protein substitution”, a transcription regulator other than Matα1 recognizes the motif adjacent to the Mcm1 site within the *C. albicans* α–specific gene *cis*-regulatory sequence. According to this model, the synthesis of this other transcription regulator would depend on Matα1, thereby preserving the regulatory logic (*14*). In the second model, “binding specificity divergence”, the binding specificity of Matα1 would have coevolved with its binding site to such an extent that the two binding sites no longer appear related by standard criterion. In the third model, the Matα1 protein would possess a relaxed specificity enabling it to recognize both *cis*-regulatory sequences.

C. albicans Matα1 activates transcription by binding to the *C. albicans* α-specific gene *cis*-regulatory sequences

To distinguish between these possibilities, we ectopically expressed the *C. albicans* Matα1 in *S. cerevisiae* MAT**a** cells (which lack *S. cerevisiae MATα1*) and assessed its ability to activate transcription from a *C. albcians* α-specific gene *cis*-regulatory sequence. We observed strong transcriptional activation by the *C. albicans* Matα1 that depended on the presence of the sequence adjacent to the Mcm1 site (Figure 2A), as well as the Mcm1 site itself (Figure S1). These results indicate that the *C. albicans* Matα1 can activate transcription by binding directly to the *C. albicans* α–specific gene *cis*-regulatory sequence. To confirm this observation, we expressed high levels of the *C. albicans* Matα1 in *S. cerevisiae* MAT**a** cells and showed by electrophoretic mobility gel shift assays on cell extracts that *C. albicans* Matα1 bound a *C. albicans* α–specific gene *cis*-regulatory sequence; incubation of the sample with a *C. albicans* Matα1 peptide antibody resulted in a super-shift (Figure 2B). Taken together, these results rule out the protein-substitution model.

Extensive DNA-binding specificity divergence of the Matα1 protein

We next addressed whether the lack of similarity between *S. cerevisiae* and *C. albicans* Matα1 binding sites reflected a true difference in the DNA-binding specificity between the two orthologs, as opposed to a relaxed Matα1 DNA-binding specificity that allows for the recognition of both sequences. We measured the ability of the *S. cerevisiae* and *C. albicans* Matα1 proteins to activate transcription from both the *S. cerevisiae* and the *C. albicans* α–specific gene *cis*-regulatory sequence and found that the Matα1 efficiently activated transcription only from the α–specific gene *cis*-regulatory sequence from its own species (Figure 3A). These findings were verified by electrophoretic gel shift assays using *S. cerevisiae* cell extracts containing either ectopically expressed *S. cerevisiae* Matα1 or *C. albicans* Matα1 (Figure 3B).

The experiments described above were performed using the *cis*-regulatory sequences from the α-mating pheromone gene, but the same results were obtained for another set of *cis*-regulatory sequences, taken from the promoters of the mating **a**-factor receptor gene (Figure S2). Additional constructs ruled out the possibility that small differences in the Mcm1 binding site could be contributing to species-specificity of Matα1 binding (Figure 3C). Taken together, these experiments lead to the conclusion that the Matα1 protein has undergone a substantial change in its DNA-binding specificity.

The DNA-binding specificity of the *C. albicans* Matα1 protein evolved after the divergence of *S. cerevisiae* and *C. albicans*

When in evolutionary history of the hemiascomycetes did the change in Matα1 DNA-binding specificity occur? To address this question, orthologs of the *S. cerevisiae* and *C. albicans* α–specific genes were identified across all available genome-sequenced yeasts. When an unambiguous ortholog could be identified, it was then determined (using PSSMs) whether a *S. cerevisiae*-like or *C. albicans*-like α–specific gene *cis*-regulatory sequence was present in the orthologous α–specific gene promoters. The *S. cerevisiae*-like *cis*-regulatory sequence appears to be present as early as the common ancestor of *S. cerevisiae* and *K. lactis* (Figure 4A), a result that was experimentally corroborated using the *K. lactis* Matα1 protein (Figure S3). Note that this analysis includes two newly sequenced fungal genomes—*Kluyveromyces wickerhamii* and *Kluyveromyces aestuarii* (see Methods & Materials).

The *C. albicans*-like sequence appears to be largely conserved across the CTG-clade (e.g.- *C. albicans, D. hansenii)*. Proceeding outward along the phylogenetic tree, we found matches to the *S. cerevisiae cis*-regulatory sequence in the filamentous fungi (e.g.- *A. terreus*, *S. sclerotiorum*), an out-group to both the *Candida* and *Saccharomyces* lineages. In fact, the filamentous fungi α–specific gene *cis*-regulatory sequence (derived from the promoters of all identifiable orthologs to either *C. albicans* or *S. cerevisaie* α–specific genes) closely resembles the *S. cerevisiae* clade α-specific gene *cis*-regulatory sequence (Figure 4B). This analysis indicates that the common ancestor to *S. cerevisiae, C. albicans,* and the filamentous fungi may have had Matα1 DNA-binding specificity similar to that of the modern *S. cerevisiae* protein and that the binding specificity of the modern *C. albicans* Matα1 changed along the evolutionary path to the common ancestor of CTG-clade. We tested this hypothesis directly by moving an α–specific gene *cis*-regulatory sequence from a filamentous fungus (*U. reesii*) into *S. cerevisiae (15)*. Expression was efficiently activated from this sequence by the *S. cerevisiae* Matα1 and only weakly activated by the *C. albicans* Matα1 (Figure 4C), consistent with the idea that the ancestral Matα1 protein possessed a *S. cerevisiae*-like DNA-binding specificity and that the most dramatic specificity change occurred in the common ancestor of the CTG-clade.

However, even within the CTG-clade, the Matα1 DNA-binding specificity did not remain constant. *C. lusitaniae* showed significant differences from the *C. albicans* in its *cis*-regulatory sequences (Figure 4A). In addition, the HMG DNA-binding domain of the *C. lusitaniae* Matα1 is the most divergent sequence among the CTG-clade Matα1 orthologs (Figure S4). To test whether these differences have consequences, we ectopically expressed *C. lusitaniae* Matα1 in *S. cerevisiae* and determined whether it could activate transcription from *cis*-regulatory sequences from either *C. lusitaniae*, *S. cerevisiae*, or *C. albicans*. Matα1 from *C. lusitiniae* efficiently activated transcription only from its own species *cis*-regulatory sequence (Figure 5B). This result indicates that Matα1 DNA-binding specificity has undergone additional changes within the CTG-clade. We also note that the α–specific gene *cis*-regulatory sequence in *Y. lipolytica* does not resemble the *C.*

albcians or *S. cerevisiae* PSSM, suggesting yet another specificity change within that lineage (Figure 4 & Figure S5).

Discussion:

We have combined bioinformatic, genetic, and biochemical experiments to demonstrate a substantial change in the DNA-binding specificity of a deeply conserved transcription regulator. Matα1 (an HMG protein) and its recognition sequence appear to have diverged substantially across the ascomycete lineage. The most dramatic changes likely occurred in the common ancestor of the CTG-clade (e.g.- *C. albicans, D. hansenii*). One manifestation of this change is that the DNA sequences recognized by Matα1 from *C. albicans* appear unrelated to those recognized by its *S. cerevisiae* ortholog. The divergence of Matα1 DNA-binding specificity is not limited to a single phylogenetic branch point, indicating that the divergence of Matα1 DNA-binding specificity has occurred multiple times.

Insights into transcription regulator DNA-binding specificity divergence

Several examples of transcription regulator DNA-binding specificity evolution have been linked to gene duplications *(16,17)*, which are hypothesized to permit drift in DNA-binding specificity by relaxing negative selection *(18)*. However, the evolution of Matα1 DNA-binding specificity demonstrates that DNA-binding specificity can extensively diverge even in the absence of gene duplication. Matα1 orthologs can be easily traced throughout the yeasts due to its conserved synteny within the MAT locus and its conserved protein sequence (Figure S5). Orthology mapping of Matα1 (see Material & Methods) across 38 genome-sequenced yeasts detected only a single, unique Matα1 ortholog in all species where the MAT locus has been sequenced. In contrast to examples of specificity changes between paralogs, Matα1 DNA-binding specificity divergence is not limited to a single phylogenetic branch point. Instead, Matα1 DNA-binding specificity appears to have diverged at several different points, indicating that DNA-binding specificity divergence between orthologous regulators can be a continuous process.

Despite this change in DNA-binding specificity, the Matα1 transcription regulator retains the same core function in both *S. cerevisiae* and *C. albicans*—activation of the α-specific genes. The conservation of function despite changes in DNA-binding specificity has been previously reported for other transcription regulators (e.g.- Rpn4 *(5)*, Yap1 *(6)*). In these cases, however, the changes in DNA-binding specificity were subtle and likely resulted from limited co-evolution of protein and DNA. We propose that the divergence of Matα1 DNA-binding specificity also represents a case of co-evolution with its recognition sequence. If so, the overall change likely occurred in stepwise fashion, perhaps the end result of numerous independent changes similar in magnitude to the DNA-binding specificity divergence between the *C. albicans* and *C. lusitaniae* Matα1. Consistent with this idea, the HMG DNA-binding domain of the *S. cerevisiae* and *C. albicans* Matα1 has undergone substantial divergence (Figure S4).

We note that most fungi have approximately five α–specific genes; although this is not a large regulon, its conserved size indicates that the evolution of the Matα1-DNA interaction occurred across a set of target genes, rather than a single gene. In addition, the interaction of Matα1 with its cofactor Mcm1 also appears to be conserved between *S. cerevisiae* and *C. albicans*. This conserved protein-protein interaction could have facilitated the evolution of Matα1 by helping to "hold it in place" while its protein-DNA interaction slowly changed.

Missing examples of DNA-binding specificity divergence

How widespread are major evolutionary changes in DNA-binding specificity by transcription regulators? There are surprisingly few documented examples of extensive DNA-binding specificity divergence between either orthologs or paralogs, a fact that has been used to argue that DNA-binding specificity evolution is uncommon in transcriptional networks. However, there is an unintended experimental bias against detecting instances of transcription regulator divergence *(19)*. There are many reasons that a regulator from one species might not function in another species; hence, these observations are rarely pursued and often left unpublished. As a result, examples of functional conservation between orthologous transcription regulators may be overrepresented in the literature *(20-22)*. For these reasons, we suggest that evolutionary changes in the DNA-binding specificity of transcriptional regulators, as documented here, may be more common than previously assumed.

The example of Matα1 DNA-binding specificity evolution has implications for bioinformatic approaches to transcriptional circuit evolution. If the only data available was the divergent *cis*-regulatory motifs, it would not be possible to distinguish between the three models described in the introduction (transcription regulator substitution, evolution of DNA-binding specificity, or relaxed DNA-binding specificity) and the observation could easily be misinterpreted. Furthermore, Matα1 DNA-binding specificity evolution demonstrates that orthologous transcription regulators can bind *cis*-regulatory sequences that appear unrelated by computational methods. This finding underscores a significant limitation of bioinformatic approaches to studying transcriptional networks that assume limited transcriptional regulator DNA-binding specificity divergence between species *(23-25)*.

Evolution of the mating-type regulatory circuitry and speciation

The evolution of Matα1 DNA-binding specificity is consistent with a network drift model of transcriptional network evolution *(26)*. In other words, the co-evolution of *cis*-regulatory sequences and transcription regulator DNA-binding specificity may have provided no specific adaptive advantage. However, it has been noted that compensatory mutations in developmental pathways, could drive speciation events through the creation of Dobhanskzy-Mueller incompatibilities *(27)*. Efficient mating in both *S. cerevisiae* and *C. albicans* requires the expression of the α-specific genes (*7,10*) and a disruption in the Matα1-DNA interaction would produce a sterile phenotype. Therefore, a mating event between an individual that had experienced Matα1/*cis*-regulatory motif compensatory evolution and an individual that had not would produce a high fraction of infertile progeny. Thus, in the absence of spatial isolation of species, coevolution of the mating regulator Matα1 and its DNA binding-sites may have contributed to speciation.

Material and Methods

PSSMs & Motif Alignments

The PSSM for the *C. albicans, K. lactis,* and *S. cerevisiae* clade α-specific gene (αsg) *cis*-regulatory sequence was derived by performing MEME *(28)* on 12, 15, and 27 sequences, respectively (see supplemental table 3 for sequence sets). The PSSM for the filamentous fungi αsg *cis*-regulatory sequence was derived by performing MEME from nine sequences identified in the promoters of αsg orthologs in the filamentous fungi species *U. reesii, C. immitis, F. graminea, A. terreus, A. nidulans,* and *S. scleotiorum (15,29).* Promoter sequences from closely-related species were pooled to increase the number of sequences submitted to MEME, thereby yielding more accurate PSSMs (under the assumption that species so closely related would not experience drastic changes in DNA-binding specificity between orthologous regulators). No close relatives of *Y. lipolytica* have been genome-sequenced *(30)*; therefore, our set of αsg orthologs for this branch was quite small (four orthologous genes). Hence, the PSSM built from six putative α-specific gene *cis*-regulatory sequences identified in *Y. lipolytica* is not as information rich as the other PSSMs presented in this work (Fig. S3). Motif alignments were computed using the motif comparison utility in MochiView *(31)*. MochiView relies on an algorithm derived from Gupta, S. et al. *(32)* to perform motif alignments. The algorithm maximizes the similarity score between two motifs and then derives an E-value from this similarity score by screening a PSSM library to determine how often this similarity score would occur by chance. The PSSM libraries that are compiled in MochiView to increase the accuracy of E-values for motif alignments are JASPAR *(33)*, SwissRegulon *(34)*, Gasch/Eisen *(5)*, Badis/Hughes *(35)*, MotifVoter *(36)*, MacIsaac *(37)*, and Zhu *(38)*.

Cloning

Primers used in this study are included in Supplemental Table S1. Due to several CUG codons in the HMG DNA-binding domain of *C. albicans MATα1*, we had the gene codon-optimized by DNA 2.0 for expression in *S. cerevisiae.* Each species *MATα1* was cloned into the 415-TEF CEN/ARS plasmid and sequenced to check for mutations *(39)*. The level of ectopic expression from these plasmids was insufficient to detect a gel-shift. Therefore, each *MATα1* was cloned into the inducible, high-expression 415-GAL 2μ plasmid *(40)*. To study αsg *cis*-regulatory sequences, 42bp regions centered-around the putative αsg *cis*-regulatory sequences for α-mating pheromone gene (except for the filamentous fungi sequence, due to the absence of a clear α-mating pheromone gene ortholog, a sequence from the promoter of mating a-factor receptor gene was used instead) were cloned into the UAS-less Cyc1 reporter construct pLG699Z *(41)* using *Xho*1. Correct orientation relative to the transcriptional start-site for the αsg *cis*-regulatory sequences within our pLG669z-derivatives was confirmed by PCR and sequencing.

Strain Construction

S. cerevisiae strains used and generated in this study are presented in Supplemental Table S2. β-galactosidase experiments were either performed in *S. cerevisiae* W303 MAT**a** cells or *S. cerevisiae* EG123 MATα Δ*matα1* strains *(42)*. Gel-shift experiments were performed using cell extracts from strains built in the *S. cerevisiae* W303 background.

β-galactosidase assays

β-galactosidase assays were performed using a standard protocol *(Rupp 2002)*. Strains were grown in SD-Ura-Lue media to maintain selection for both plasmids. For each strain five colonies were grown overnight, diluted back, and allowed to reach log phase. Cells were harvested, permeablized, and activation assays performed. Data in any figure panel are from the same day.

Electrophoretic Mobility Shift Assays

Yeast strains were grown overnight night in either glucose or galactose medium (in both media-types selection was maintained for the plasmid marker), depending on whether ectopic expression of Matα1 was desired. Harvested cells were of an OD_{600} between 0.75 and 1.0. *S. cerevisiae* pellets were resuspended in 100 mM Tris [pH = 8], 200 mM NaCl, 1 mM EDTA, 10 mM $MgCl_2$, 10 mM β-mercapoethanol, 20% glycerol, and Roche Complete protease inhibitors (1 tablet/10 ml). Extracts were lysed by sonification and then cleared by centrifugation at 12,000 x g for 20 minutes, yielding ~10 mg/ml of total protein. Electrophoretic mobility gel shift assays were performed using *S. cerevisiae* cell extracts as described in Keleher, C.A. et al. *(43)*. αsg *cis*-regulatory sequence oligonucleotide probes were labeled with P^{32} γ-ATP using T4 PNK. Binding conditions were 50 mM Tris [pH = 8], 100 mM NaCl, 10% Glycerol, 5 mM MgCl2, 5mM β-mercapoethanol, 50μg/mL Poly(dI-dC) (limits non-specific protein:DNA-binding), and 1.2 μM labeled oligonucleotide. Antibody supershifts were accomplished using a Matα1 N-terminal peptide antibody (Bethyl Antibodies, antigenic sequence—MGNKKKTRKTVPKEFISLC). For a 20 μl Protein:DNA-binding reaction, 0.5 μl of a 1:100 dilution of immune serum was sufficient to induce supershifts.

Orthology Mapping

Orthology mapping was performed as described in Tsong et al. *(44)*. *S. cerevisiae* and *C. albicans* αsg protein sequences were used to "query" a single database containing all ORF sequences from 38 fungal species using PSI-BLAST *(45)*, employing an E-value cutoff of 10^{-5} and the Smith-Waterman alignment option. The sequences returned by PSI-BLAST were then multiply-aligned with ClustalW (using the fast alignment option) and a neighbor joining tree (NJ) was inferred, again using ClustalW *(46)*. Finally, the resulting NJ tree was traversed to extract a set of orthologous genes.

Genome Sequencing

To improve our ability to detect *cis*-regulatory sequences in the emerging yeast model organism *Kluyveromyces lactis* using phylogenetic footprinting *(47)*, the genomes of the two close relatives of the *Kluyveromyces lactis (Kluyveromyces aestuarii* (ATCC 18862) and *Kluyveromyces wickerhamii* (UCD 54-210)) were sequenced. *Kluyveromyces aestuarii* was sequenced to an estimated coverage of 14X and *Kluyveromyces wickerhamii* to an estimated coverage of 12X coverage on a 454 platform at the Washington University Genome Sequencing Center. The Washington University Genome Sequencing Center used the assembly algorithm Newbler in early 2008 to assemble the 454 reads into contigs. This level of sequencing was insufficient to assemble complete chromosomes, but was sufficient to extract information about αsg orthologs in these species. For *Kluyveromyces wickerhamii,* after assembly, the number of long contigs (>500bp) was 510 and the number of short contigs (>100bp) was 953. For *Kluyveromyces aestuarii*, the number of long contigs (>500bp) was 336 and the number of short contigs (>100bp) was 682. The sequence will be available through the Johnson lab website, along with ORF calls, and is currently available through GenBank as a whole-genome shotgun sequencing project data (Genbank numbers: *K. aestuarii*- AEAS00000000 & *K. wickerhamii*- AEAV00000000).

Acknowledgements:

Work of the authors was supported by grants from the National Institutes of Health (AI49187) and the Ellison Foundation (ID-SS-0628-04). The authors thank Oliver Homann and Xin He for sharing their knowledge of bioinformatics. The authors also thank Oliver Homann and Linet Mera for valuable comments on the manuscript.

References

1. Wray, G. A. et al. (2003) The Evolution of Transcriptional Regulation in Eukaryotes. *Mol Biol Evol* 20:1377-1419.
2. Prud'homme, B., Gompel, N., Carroll, S.B. (2007) Emerging Principles of Regulatory Evolution. *Proc Natl Acad Sci U S A* 104:8605-8612.
3. McGinnis, N., Kuziora, M. A. & McGinnis, W. Human (1990) Hox-4.2 and Drosophila deformed encode similar regulatory specificities in Drosophila embryos and larvae. *Cell* 63:969-976.
4. Halder, G., Callaerts, P. & Gehring, W. J. (1995) Induction of ectopic eyes by targeted expression of the eyeless gene in Drosophila. *Science* 267**:**1788-1792.
5. Gasch, A. P. et al. (2004) Conservation and evolution of cis-regulatory systems in ascomycete fungi. *PLoS Biol* 2:e398.
6. Kuo, D. et al. (2010) Coevolution within a transcriptional network by compensatory *trans* and *cis* mutations. *Genome Res* 20:1672-1678.

7. Bender, A. & Sprague, G. F., Jr. (1987) MAT alpha 1 protein, a yeast transcription activator, binds synergistically with a second protein to a set of cell-type-specific genes. *Cell* 50:681-691.

8. Jarvis, E.E., Clark, K.L., & Sprague, G.F. (1989) The yeast transcription activator PRTF, a homolog of the mammalian serum response factor, is encoded by the Mcm1 gene. *Gen & Dev* 3:936-945.

9. Taylor, J. W. & Berbee, M. L. (2006) Dating divergences in the Fungal Tree of Life: review and new analyses. *Mycologia* 98:838-849.

10. Tsong, A. E., Miller, M. G., Raisner, R. M. & Johnson, A. D. (2003) Evolution of a combinatorial transcriptional circuit: a case study in yeasts. *Cell* 115:389-399.

11. Tuch, B. B., Galgoczy, D. J., Hernday, A. D., Li, H. & Johnson, A. D. (2008) The evolution of combinatorial gene regulation in fungi. *PLoS Biol* 6:e38.

12. Kellis M. et al. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423:241-254.

13. Butler, G. et al. (2009) Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. *Nature* 459:657-662.

14. Booth, L.N., Tuch, B.T., & Johnson, A.D. (2010) Intercalation of a new tier of transcription regulation into an ancient genetic circuit. *Nature* 468:959-963.

15. Sharpton, T. J. et al. (2009) Comparative genomic analyses of the human fungal pathogens *Coccidioides* and their relatives. *Genome Res* 19:1722-1731.

16. Wharton, R. P. & Ptashne, M. (1985) Changing the binding specificity of a repressor by redesigning an alpha-helix. *Nature* 316:601-605.

17. Knight, K. L. & Sauer, R. T. (1989) DNA-binding specificity of the Arc and Mnt repressors is determined by a short region of N-terminal residues. *Proc Natl Acad Sci USA* 86:797-801.

18. Emerson, R. O. & Thomas, J. H. (2009) Adaptive Evolution in Zinc Finger Transcription Factors. *PLoS Gen* 5:e1000325.

19. Lynch, V. J. & Wagner, G. P. (2008) Resurrecting the role of transcription factor change in developmental evolution. *Evolution* 62:2131-2154.

20. Ranganayakulu, G. et al. (1998) Divergent roles for NK-2 class homeobox genes in cardiogenesis in flies and mice. *Development* 125:3037–3048.

21. Park, M. et al. (1998) Differential rescue of visceral and cardiac defects in Drosophila by vertebrate tinman-related genes. *Proc Natl Acad Sci USA* 95:9366–9371.

22. Maizel, A. et al. (2005) The floral regulator LEAFY evolves by substitutions in the DNA-binding domain. *Science* 308:260-263.

23. Xie, X. et al. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* 434:338-345.

24. Doninger, S.W., Fay, J.C. (2007) Frequent gain and loss of functional transcription factor binding sites. *PLoS Comput Biol* 3:e99.

25. Li, X.Y. et al. (2008) Transcription factors bind thousand of active and inactive regions in the *Drosophila* blastoderm. *PLoS Biol* 6:e27.

26. Lynch, M. (2007) The Frailty of adaptive hypotheses for the origins of organismal complexity. *Proc Natl Acad Sci U S A* 104:8597-8604.

27. Porter, A.H. & Johnson, N.A. (2002) Speciation despite gene flow when developmental pathways evolve. *Evolution* 56:2103-2111.

28. Bailey, T. A. & Elkan, C. (1994) "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28-36, AAAI Press, Menlo Park, California.

29. Dietrich, F. S. et al. (2004) The *Ashbya gossypii* genome as a tool for mapping the ancient *Sacchoaromyces cerevisiae* genome. *Science* 304:304-207.

30. Dujon, B. et al. (2004) Genome evolution in yeasts. *Nature* 430:35-44.

31. Homann, O. R. & Johnson, A. D. (2010) MochiView: versatile software for genome browsing and DNA motif analysis. BMC Biol 8:49.

32. Gupta, S. et al. (2007) Quantifying similarity between motifs. *Genome Biol* 8:R24.

33. Sandelin, A. et al. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* 32:D91-D94.

34. Pachkov, M. et al. (2006) SwissRegulon: a database of genome-wide annotations of regulatory sites. *Nucleic Acids Res* 27:D1-D5.

35. Badis, G. et al. (2009) Diversity and Complexity in DNA recognition by Transcription Factors. *Science* 324:1720-1723.

36. Wijaya, E. et al. (2008) MotifVoter: a novel ensemble method for fine-grained integration of generic motif finder. *Nucleic Acids Res* 24:2288-2295.

37. MacIsaac, K.D. et al. (2006) An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics* 7:7:113.

38. Zhu, C. et al. (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res* 4:556-566.

39. Mumberg, D., Müller, R., Funk M. (1995) Yeast vectors for the controlled expression of heterologous proteins in different genetic backgrounds. *Gene* 156:119-122.

40. Mumberg, D., Müller, R., Funk M. 8 (1994) Regulatable promoters of *Saccharomyces cerevisiae*: comparison of transcriptional activity and their use for heterologous expression. *Nucleic Acids Res* 22:5767-576.

41. Guarente, L. & Ptashne, M. (1981) Fusion of Escherichia coli lacZ to the cytochrome c gene of Saccharomyces cerevisiae. *Proc Natl Acad Sci USA* 78:2199-2203.

42. Galgoczy, D. J. et al. (2004) Genomic dissection of the cell-type-specification circuit in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA* 101:18069-18074.

43. Keheler, C.A. Passmore, S., Johnson, A.D. (1989) Yeast repressor alpha 2 binds cooperatively with yeast protein Mcm1. *Mol Cell Biol* 9:5228-5230.

44. Tsong, A.E. et al. (2006) Evolution of alternative transcriptional circuits with identical logic. *Nature* 443:415-420.

45. Altschul, S. F. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389-402.

46. Higgins, D. G. & Sharp, P. M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* 73: 237-44.

47. Cliften, P. et al. (2003) Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* 301:71-76.

48. Scannell, D.R. et al. (2007) Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. *P.N.A.S. USA* 104:8397-8402.

Figure 1

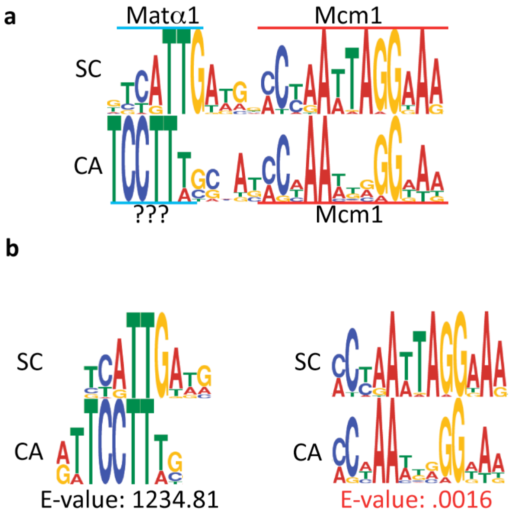


Figure 1: Significant divergence of the α-specific gene *cis*-regulatory sequence between *C. albicans* and *S. cerevisiae*. (a) The PSSM for the *S. cerevisiae* clade α-specific gene (αsg) *cis*-regulatory sequence (SC) was derived using MEME from 27 sequences identified in either the promoters of known *S. cerevisiae* αsgs *(42)* or in promoters of the orthologous genes in *S. mikatae, S. paradoxus,* and *S. bayanus*. The PSSM for the *C. albicans* clade αsg *cis*-regulatory sequence (CA) was derived using MEME from 12 sequences that originated from either *C. albicans* αsg promoter sequences *(10)* or from the promoters of the orthologous

genes in *C. tropicalis* and *C. dublienesis*. (b) Alignments of the *S. cerevisiae* Matα1 motif to the unknown motif within the *C. albicans* αsg *cis*-regulatory sequence (left) and the αsg Mcm1 motif from *S. cerevisiae* and *C. albicans* (right). Motif alignments and E-values were calculated using MochiView *(30)*, which quantifies similarities between motifs by using an algorithm derived from Gupta, S. et al. *(31)*.

Figure 2

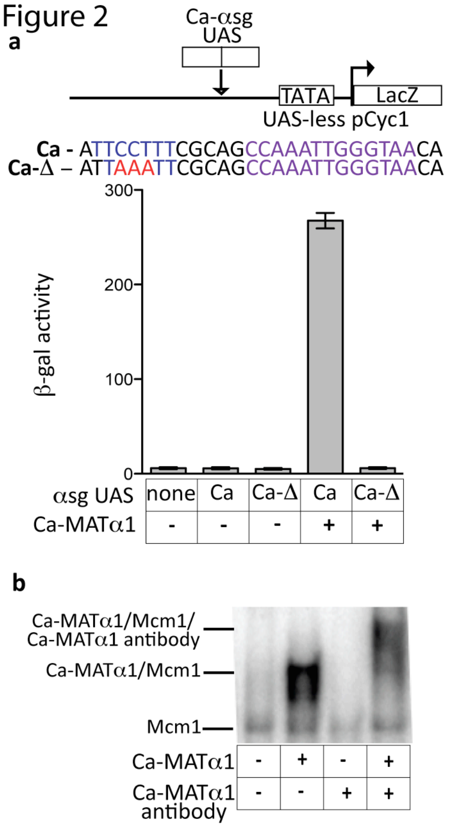


Figure 2: *C. albicans* Matα1 activates transcription by binding to the *C. albicans* α-specific gene *cis*-regulatory sequences. (a) A *C. albicans* αsg *cis*-regulatory sequence taken from the α-mating pheromone gene was inserted into a basal

promoter construct upstream of a β-gal reporter (pLG669Z). The same *C. albicans* αsg *cis*-regulatory sequence was also mutated to alter the residues at the position where Matα1 binds to the *S. cerevisiae cis*-regulatory sequence (Ca-Δ). These constructs were introduced into *S. cerevisiae* MAT**a** cells (MAT**a** cells lack *S. cerevisiae MATα1*). In the two right lanes, strains also contain a 415-TEF plasmid modified to express a codon-changed *C. albicans* Matα1 (the codon changes were necessary because *C. albicans* decodes the CUG codon as serine and most other species, including *S. cerevisiae*, decode it as a luecine). Reporter activity was monitored using β-galactosidase assays. Each sample has an N = 5 and error bars represent standard error. (b) Electophoretic mobility gel shift assays were performed using *S. cerevisiae* cell extracts. The labeled oligonucleotide used in this experiment was the *C. albicans* αsg *cis*-regulatory sequence described in part A. Extracts were prepared from a *S. cerevisiae* MATa strain containing a galactose-inducible copy of the codon-changed *C. albicans* Matα1. Each lane contains 5 mg of protein from cell extracts. Galactose induction was performed overnight on samples in lanes 2 and 4 (lanes 1 and 3 are grown in glucose, turning off *C. albicans* Matα1 expression). In lanes 3 and 4, a N-terminal peptide antibody against *C. albicans* Matα1 (Bethyl Laboratories) was used to confirm that DNA-binding activity was due to the *C. albicans* Matα1 protein.

Figure 3

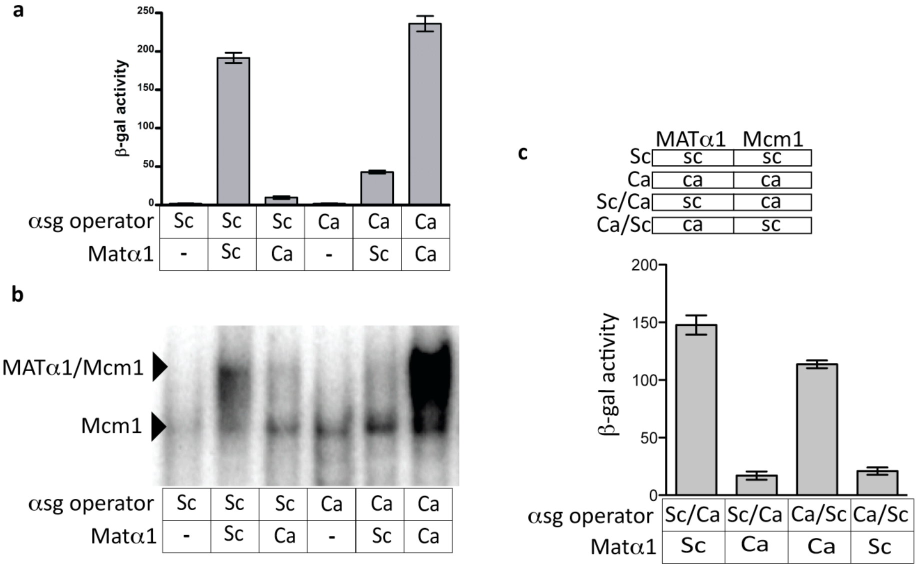
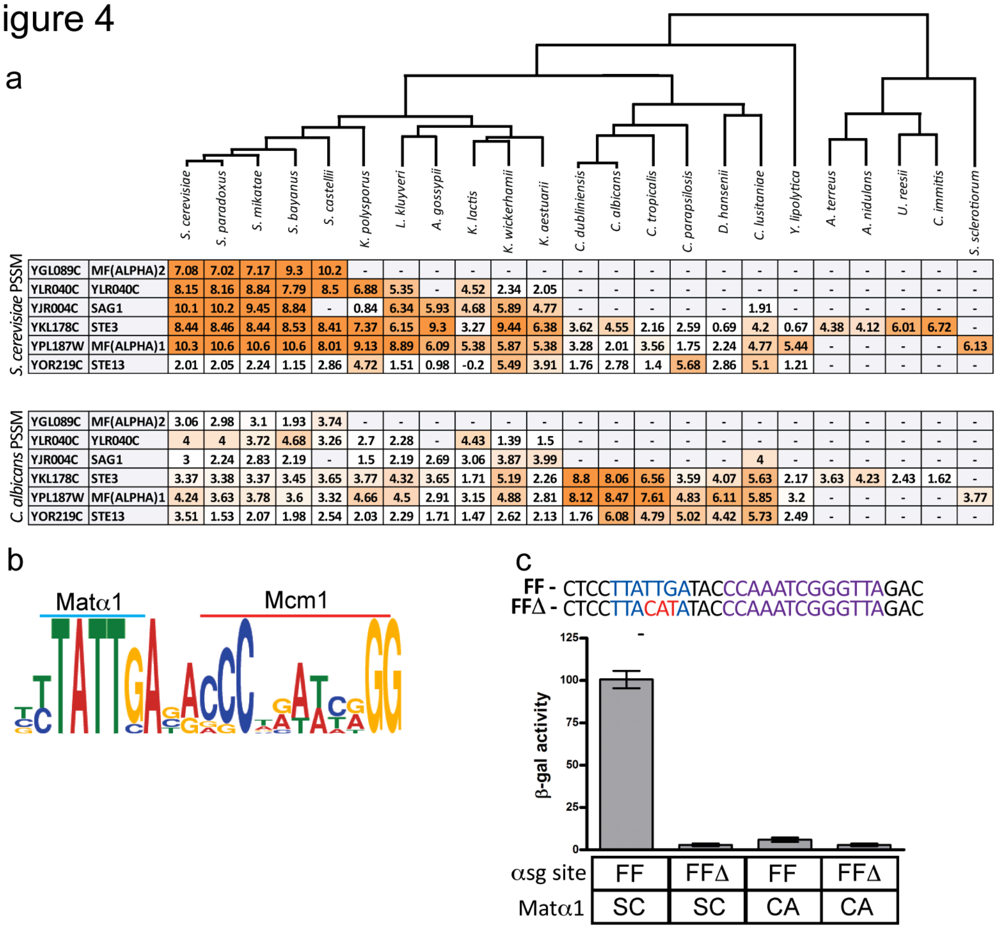


Figure 3- Extensive DNA-binding specificity divergence of the Matα1 protein. (a) The αsg *cis*-regulatory sequence of the promoter for the α-mating pheromone from *C. ablicans* (Ca) or from *S. cerevisiae* (Sc) was inserted into a basal promoter construct (pLG669z). These constructs were introduced into *S. cerevisiae* MATα Δ*matα1* cells along with a 415 TEF plasmid modified to express *S. cerevisiae MATα1* (lanes 2 & 5) or a 415 TEF plasmid modified to express the codon-changed *C. albicans MATα1* (lanes 3 & 6). Reporter activity was monitored using β-galactosidase assays. Each sample has an N = 5 and error bars represent standard error. (b) Electrophoretic mobility gel shift assays were performed using *S. cerevisiae* cell extracts. The labeled oligonucleotide used in this experiment was either the *C. albicans* αsg *cis*-regulatory sequence (lanes 4-6) or *S. cerevisiae* αsg *cis*-regulatory sequence (lanes 1-3), both described in part A. Extracts were prepared from either *S. cerevisiae* MAT**a** cells containing a galactose-inducible copy of *C. albicans MATα1* or *S. cerevisiae* MATα cells

containing a galactose-inducible copy of the *S. cerevisiae MATα1* (p415GAL). Galactose induction was performed overnight on samples in lanes 2, 3, 5, & 6 (lanes 1 & 4 are grown in glucose). Each lane contains 5 mg of protein from cell extracts. (c) To create the Ca/Sc hybrid construct, the Matα1 binding site from the Ca reporter construct was used to replace the Matα1 binding site in the Sc reporter construct. To create the Sc/Ca hybrid construct, the Matα1 binding site from the Sc reporter construct was used to replace the Matα1 binding site in the Ca reporter construct. Reporter activity was monitored using β-galactosidase assays.

Figure 4



S. cerevisiae PSSM			S. cerevisiae	S. paradoxus	S. mikatae	S. bayanus	S. castellii	K. polysporus	L. kluyveri	A. gossypii	K. lactis	K. wickerhamii	K. aestuarii	C. dubliniensis	C. albicans	C. tropicalis	C. parapsilosis	D. hansenii	C. lusitaniae	Y. lipolytica	A. terreus	A. nidulans	U. reesii	C. immitis	S. sclerotiorum
	YGL089C	MF(ALPHA)2	7.08	7.02	7.17	9.3	10.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	YLR040C	YLR040C	8.15	8.16	8.84	7.79	8.5	6.88	5.35	-	4.52	2.34	2.05	-	-	-	-	-	-	-	-	-	-	-	-
	YJR004C	SAG1	10.1	10.2	9.45	8.84	-	0.84	6.34	5.93	4.68	5.89	4.77	-	-	-	-	-	1.91	-	-	-	-	-	-
	YKL178C	STE3	8.44	8.46	8.44	8.53	8.41	7.37	6.15	9.3	3.27	9.44	6.38	3.62	4.55	2.16	2.59	0.69	4.2	0.67	4.38	4.12	6.01	6.72	-
	YPL187W	MF(ALPHA)1	10.3	10.6	10.6	10.6	8.01	9.13	8.89	6.09	5.38	5.87	5.38	3.28	2.01	3.56	1.75	2.24	4.77	5.44	-	-	-	-	6.13
	YOR219C	STE13	2.01	2.05	2.24	1.15	2.86	4.72	1.51	0.98	-0.2	5.49	3.91	1.76	2.78	1.4	5.68	2.86	5.1	1.21	-	-	-	-	-

C. albicans PSSM																									
	YGL089C	MF(ALPHA)2	3.06	2.98	3.1	1.93	3.74	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	YLR040C	YLR040C	4	4	3.72	4.68	3.26	2.7	2.28	-	4.43	1.39	1.5	-	-	-	-	-	-	-	-	-	-	-	-
	YJR004C	SAG1	3	2.24	2.83	2.19	-	1.5	2.19	2.69	3.06	3.87	3.99	-	-	-	-	-	4	-	-	-	-	-	-
	YKL178C	STE3	3.37	3.38	3.37	3.45	3.65	3.77	4.32	3.65	1.71	5.19	2.26	8.8	8.06	6.56	3.59	4.07	5.63	2.17	3.63	4.23	2.43	1.62	-
	YPL187W	MF(ALPHA)1	4.24	3.63	3.78	3.6	3.32	4.66	4.5	2.91	3.15	4.88	2.81	8.12	8.47	7.61	4.83	6.11	5.85	3.2	-	-	-	-	3.77
	YOR219C	STE13	3.51	1.53	2.07	1.98	2.54	2.03	2.29	1.71	1.47	2.62	2.13	1.76	6.08	4.79	5.02	4.42	5.73	2.49	-	-	-	-	-

Figure 4- The DNA-binding specificity of the *C. albicans* Matα1 protein evolved after the divergence of *S. cerevisiae* and *C. albicans.* (a) Orthologs of the *S. cerevisiae* and *C. albicans* αsgs were mapped across 38-genome sequenced yeasts *(10,11,13,28,46,48)*. Where a clear ortholog could be detected, the promoters of these orthologs were scanned with either the *S. cerevisiae* or *C. albicans* clade αsg *cis*-regulatory sequence PSSM (created as described in Figure legend 1A). Maximum $\log_{10}$ odds-scores are shown. Darker shades of orange indicate a stronger match to the PSSM. One-to-one orthologs become more difficult to detect with greater evolution distance, hence the small number of orthologs identified in the filamentous fungi. (b) The PSSM for the filamentous fungi αsg *cis*-regulatory sequence was derived using MEME from nine sequences identified in the promoters of αsg orthologs in the filamentous fungi species *U. reesii, C. immitis, F. graminea, A. terreus, A. nidulans,* and *S. scleotiorum*. (c) A putative αsg *cis*-regulatory sequence from the promoter of the *STE3* ortholog in the filamentous fungi species *U. reesii* (FF) was placed into the basal promoter construct (pLG669z). The same construct was mutated at the position of the putative Matα1 motif (FF-Δ). *S. cerevisiae* Matα1 was supplied by the endogenous copy within a MATα strain (lanes 1 & 2) and *C. albicans* Matα1 from expression off p415TEF within a MAT**a** strain (lanes 3 &4). Reporter activity was monitored using β-galactosidase assays. Each sample has an N = 5 and error bars represent standard error.

Figure 5

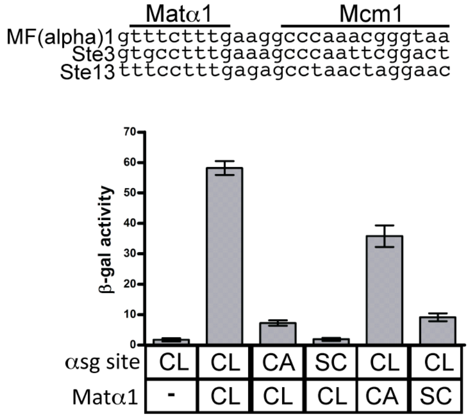
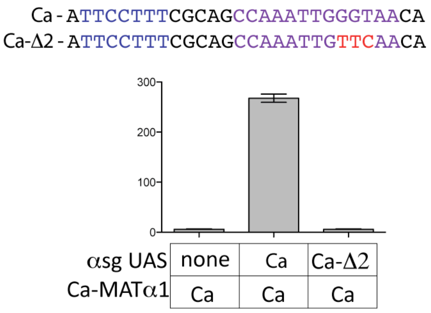


Figure 5- Matα1 DNA-binding specificity has continued to diverge within the CTG-clade. (a) Three putative αsg *cis*-regulatory sequences identified by MEME in the promoters of *C. lusitaniae* αsg orthologs. (b) The αsg *cis*-regulatory sequence of the promoter for the α-mating pheromone (*MFα1*) from *C. lusitaniae* (Cl) was inserted into a basal promoter construct (pLG669z) and the *C. lusitaniae* Matα1 was expressed from a 415 TEF plasmid. Plasmids were transformed into a *S. cerevisiae* MATα Δ*matα1* strain. Reporter activity was monitored using β-galactosidase assays. Each sample has an N = 5 and error bars represent standard error.

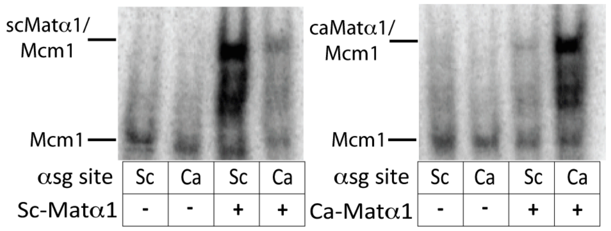
Supplemental Figure 1



Supplemental Figure Legend 1: The Mcm1 binding site within *C. albicans* αsg *cis*-regulatory sequence is necessary for Ca-Matα1-dependent transcriptional activation

In *S. cerevisiae*, Matα1 synergistically binds αsg *cis*-regulatory sequence with Mcm1 *(8)*. Disruption of the Mcm1 binding site causes a complete loss in Matα1-dependent transcriptional activation *(8)*. The co-occurrence of Matα1 and Mcm1 binding sites at a fixed distance in *C. albicans* strongly suggests that Matα1 and Mcm1 are also acting synergistically at *C. albicans* αsg *cis*-regulatory sequences. To test whether, like *S. cerevisiae*, the Mcm1 binding site within *C. albicans* αsg *cis*-regulatory sequence was essential to Matα1-dependent transcriptional activation, we mutated the Mcm1 binding site within a *C. albicans* αsg *cis*-regulatory sequence promoter construct (Ca) to create the Ca-Δ2 reporter construct. Activity was monitored by β-galactosidase assays. Consistent with a synergeristic interaction between Mcm1 and *C. albicans* Matα1, disruption of the Mcm1 binding site eliminated all *C. albicans* Matα1-dependent transcriptional activation.

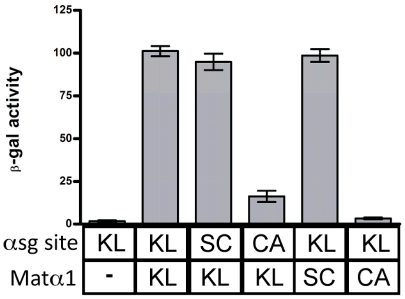
Supplemental Figure 2



Supplemental Figure Legend 2: Species-specific binding to a second set of αsg-*cis*-regulatory sequences

In the reporter construct and gel-shift experiments, the αsg *cis*-regulatory sequences used were derived from the *C. albicans* and *S. cerevisiae* α-mating pheromone *cis*-regulatory sequences. Thus, it was a concern that the results were dependent on this specific set of αsg *cis*-regulatory sequences. To directly address this concern, electrophoretic mobility gel shift assays were performed on a second set of αsg *cis*-regulatory sequences taken from the promoter sequences for the *C. albicans* and *S. cerevisiae* **a**-factor receptor genes (*STE3*). The labeled oligonucleotide used in this experiment was either the *C. albicans STE3* αsg *cis*-regulatory sequence (lanes 2 & 4) or *S. cerevisiae STE3* αsg *cis*-regulatory sequence (lanes 1 & 3) in both images. Extracts were prepared from either *S. cerevisiae* MAT**a** strain containing a galactose inducible copy of *C. albicans MATα1* or *S. cerevisiae* MATα cells containing a galactose-inducible copy of the *S. cerevisiae MATα1* (p415GAL). Galactose induction was performed overnight on samples in lanes 3 & 4 (lanes 1 & 2 are grown in glucose) in both images. Each lane contains 5 mg of protein from cell extracts.

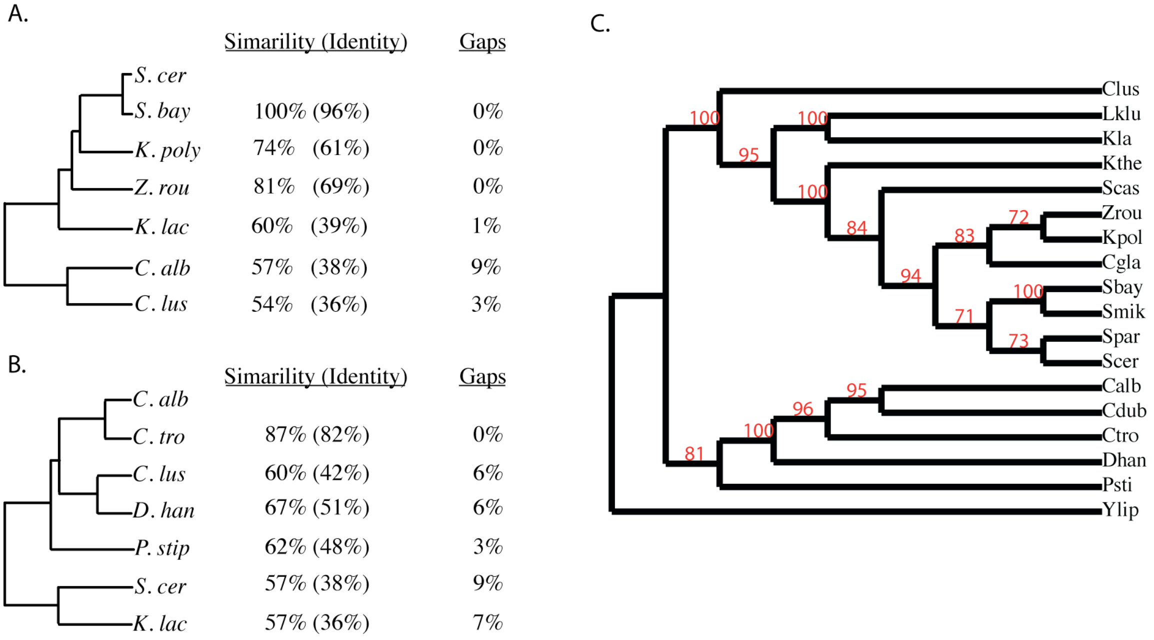
Supplemental Figure 3



Supplemental Figure Legend 3: Functional conservation of Matα1 DNA binding-specificity between *S. cerevisiae* and *K. lactis*

The αsg *cis*-regulatory sequence of the promoter for the α-mating pheromone gene from *K. lactis* (Kl) was inserted into a basal promoter construct (pLG669z) and the *K. lactis* Matα1 was expressed from a 415 TEF plasmid. Plasmids were transformed into a *S. cerevisiae* MATα Δ*mata1* strain. Reporter activity was monitored using β-galactosidase assays.

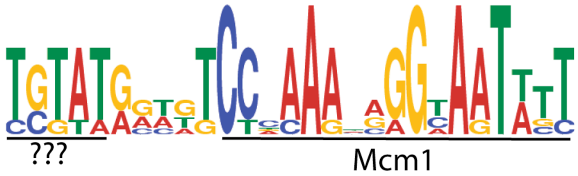
Supplemental Figure 4



Supplemental Figure Legend 4: Evolution of the Matα1 HMG DNA-binding domain

A multiple protein sequence alignment for the Matα1 HMG DNA-binding domain was computed using ClustalW2. (A) Quantification of Matα1 HMG DNA-binding domain divergence relative to the *S. cerevisiae* Matα1 sequence. Percent protein sequence similarity, identity, and gaps were calculated using BLAST2. (B) Quantification of Matα1 HMG DNA-binding domain divergence relative to the *C. albicans* Matα1 sequence. Note that the *C. lusitaniae* and *S. cerevisaie* Matα1 sequences have comparable similarity scores to *C. albicans* Matα1 despite their extensive difference in phylogenetic proximity to *C. albicans.* (C) The *C. lusitaniae* (Clus) Matα1 sequence does not branch with the other CTG-clade species in a tree constructed using the Matα1 HMG DNA-binding domain. The tree was generated using the ClustalW2 alignment of the Matα1 HMG DNA-binding domain. The bootstrap values in support of this particular branching configuration are shown in red. Notably, in contrast to the ascomycete phylogenetic tree, the *C. lusitaniae* Matα1 branches outside the CTG-clade. This result is consistent with the divergent DNA-binding specificity of *C. lusitaniae* relative to *C. albicans.*

Supplemental Figure 5



Supplemental Figure Legend 5: *Y. lipolytica* αsg *cis*-regulatory PSSM

In Figure 4, neither the *S. cerevisiae* nor the *C. albicans* PSSMs identified close-matches in the promoters of *Y. lipolytica* αsgs. Utilizing MEME, six sequences were located in the promoters of *Y. lipolytica* αsgs. The PSSM built from these six sequences contains a Mcm1 binding site and corroborating the results of Figure 4, the sequence at the position of the Matα1 binding site does not resemble either the *C. albicans* or the *S. cerevisiae* Matα1 site. Unfortunately, no genomes have been sequenced for species closely-related to *Y. lipolytica*, which could provide further support for the possibility of yet another change in Matα1 DNA-binding specificity.

Chapter 3

Protein modularity, cooperative binding, and hybrid regulatory states underlie transcriptional network diversification

Christopher R. Baker*, Lauren N. Booth*, Trevor R. Sorrells, and Alexander D. Johnson

Department of Microbiology and Immunology and Department of Biochemistry
The University of California, San Francisco
San Francisco, CA 94102
USA

*These authors contributed equally and are listed alphabetically

Summary

We examine how different transcriptional network structures can evolve from an ancestral network. By characterizing how the ancestral mode of gene regulation for genes specific to **a**-type cells in yeast species evolved from an activating paradigm to a repressing one, we show that regulatory protein modularity, conversion of one *cis*-regulatory sequence to another, distribution of binding energy among protein-protein and protein-DNA interactions, and exploitation of ancestral network features all contribute to the evolution of a novel regulatory mode. The formation of this derived mode of regulation did not disrupt the ancestral mode and thereby created a hybrid regulatory state where both means of transcription regulation (ancestral and derived) contribute to the conserved expression pattern of the network. Finally, we show how this hybrid regulatory state has resolved in different ways in different lineages to generate the diversity of regulatory network structures observed in modern species.

Research Highlights

- Protein modularity & ancestral feature exploitation bypass evolutionary constraint
- Gain of new regulator-regulator interaction transformed a transcription network
- This gain resulted in a hybrid state with ancestral & derived regulatory features
- Partial redundancy of the hybrid state enables regulatory network diversification

Introduction

In many organisms, gene regulatory networks have been shown to undergo significant divergence over evolutionary time (reviewed by (Doebley and Lukens 1998; Carroll 2005; Davidson and Erwin 2006; Wray 2007; Tuch, Li et al. 2008; Wohlbach, Thompson et al. 2009). In the simplest cases, the gain or loss of a *cis*-regulatory sequence upstream of a single gene can produce changes in coloration, losses of ancestral anatomical features, or altered ability to digest sugars (Gompel, Prud'homme et al. 2005; Tishkoff, Reed et al. 2007; Chan, Marks et al. 2010). Yet, it seems likely that the evolution of complex biological innovations requires concerted evolution across entire networks of genes (Tuch, Li et al. 2008; Lavoie, Hogues et al. 2010; Lynch, Leclerc et al. 2011). Two considerations suggest that network evolution requires mechanisms in addition to the loss and gain of single *cis*-regulatory sequences. First, the adaptive value of acquiring coordinated expression of a large set of genes may not be realized until all or at least a large fraction of the gene set acquires the new regulatory input. Second, expression of only a portion of the gene network could be detrimental to the fitness of the organism, for example, through the non-stoichiometric expression of components of a protein complex.

To understand the molecular events that underlie changes in the regulation of groups of genes, we investigated a transcriptional network that determines cell-type in a wide variety of fungal species. This network—comprised of the **a**-specific genes (**a**sgs) and their regulators—underwent a major circuit rewiring in the hemiascomycete yeasts (Tsong, Miller et al. 2003; Tsong, Tuch et al. 2006). This group of yeast includes *Saccharomyces cerevisiae* (the baker's yeast), *Kluyveromyces lactis* (a dairy yeast), *Candida albicans* (the most common human fungal pathogen), and over 30 additional genome-sequenced species (Figure 1A). This lineage has been estimated to represent at least 300 million years of evolutionary time (Taylor and Berbee 2006). Virtually all of yeast species in the hemiascomycete lineage exist in three cell types—the mating competent **a** and α cells and the product of their mating, the **a**/α cell (Figure 1B). Mating cell-type is controlled by transcriptional regulators that are encoded at the mating-type (*MAT*) locus (Herskowitz 1989). These regulators control the expression of genes that are responsible for the specialized properties of each of the three cell types. The **a**sgs are a group of seven to ten genes (depending on the species) whose key regulatory characteristic is that they are expressed in the **a** cell-type but not in the α and **a**/α cell-types (Herskowitz 1989; Tsong, Miller et al. 2003; Galgoczy, Cassidy-Stone et al. 2004) (Figure 1B). The **a**sgs encode proteins (e.g. α mating pheromone receptor, **a** mating pheromone, agglutinins and exporters) that are necessary for the specific properties of **a** cells (Herskowitz 1989) (Madhani 2007).

In principle, there are two ways that the **a**sgs could be expressed in **a** cells but not in the other two cell types: (1) the **a**sgs could be activated by a regulatory protein present only in **a** cells or (2) the **a**sgs could be repressed by a regulator made only in α and **a**/α cells. In fact, both schemes are observed, the latter in *S. cerevisiae* and the former in *C. albicans* and (Strathern, Hicks et al. 1981; Tsong, Miller et al. 2003). In *C. albicans*, the HMG domain protein **a**2 binds to and activates the **a**sgs. In *S. cerevisiae*, the homeodomain protein α2 binds to and represses the **a**sgs (Johnson and Herskowitz 1985). We previously showed that the activation mode of regulation (by **a**2) was present in the ancestor of *C. albicans* and *S. cerevisiae* and that the switch to the repression mode (mediated by α2) occurred along the branch to *S. cerevisiae* (Tsong, Tuch et al. 2006). Indeed, the gene encoding the **a**2 protein was lost from the genome in an ancestor of *S. cerevisiae* (Butler et al., 2004). (Figure 1C)

Here we define the evolutionary path for the switch in regulation of the **a**sg network using a combination of bioinformatic analysis, direct experiments in the yeasts *Kluyveromyces wickerhamii, Kluyveromyces lactis*, and *Lachancea kluyveri,* ancestral protein reconstruction, and *trans*-species reporter gene analysis in *S. cerevisiae*. Our principle conclusions are as follows: First, regulatory protein modularity was crucial for the change in network regulation. In particular, protein modularity accounts for the cooption of an existing repressor for a new function (repression of the **a**sgs) while maintaining its ancestral function. Second, the cooperative binding of transcriptional regulators facilitated the gain of the repression mode of regulation across this gene set by stabilizing early evolutionary intermediates. Third, the conversion of one *cis*-regulatory sequence into another occurred through an "intermediate" *cis*-regulatory sequence that was recognized by regulators of both the ancestral and derived regulatory modes. Fourth, the evolution of **a**sg repression in the common ancestor of *K. lactis* and *S. cerevisiae* did not disrupt the ancestral (positive) mode of regulation, and thereby formed a "hybrid" regulatory state (Tsong, Tuch et al. 2006). Finally, we show that once the hybrid regulatory network formed, it resolved in different ways along the branches to the modern yeast species: in *S. cerevisiae* the ancestral form was discarded, leaving only the derived form; in *K. lactis* the derived form was inactivated, reverting to the ancestral mode of regulation; in *L. kluyveri* and *K. wickerhamii,* aspects of the hybrid regulatory state have been maintained. Because the regulatory proteins studied here are conserved in all eukaryotes, the evolution of **a**sg regulation can serve as a model for understanding the molecular mechanisms underlying the extraordinary flexibility of transcriptional circuits over evolutionary time.

Results

α2 repression of the **a***sgs evolved prior to the divergence of Saccharomyces and Kluyveromyces*

We determined the time at which repression of the **a**sgs arose during evolutionary time. To do this, we moved the **a**sg regulatory sequences (from the conserved **a**sg *STE2*) and the α2 proteins from a variety of species into *S. cerevisiae* and determined their abilities to support repression (Fig. 2A). In *S. cerevisiae*, α2 binds **a**sg *cis*-regulatory sequences cooperatively with a MADS-box transcription regulator, Mcm1 (Figure 1C). Both proteins bind with high affinity to DNA sequences and their cooperative binding results from a relative weak protein-protein interaction (Vershon and Johnson 1993; Tan and Richmond 1998). The *cis*-regulatory sequence consists of an Mcm1 homodimer site flanked by two α2 binding sites (Keleher, Goutte et al. 1988). Removal of any these four binding sites from an **a**-specific *cis*-regulatory sequence, or disruption of the protein-protein interaction, severely compromises repression (Vershon and Johnson 1993; Smith and Johnson 1994).

The *STE2 cis*-regulatory sequences from species that branch from the *S. cerevisiae* lineage prior to the loss of the **a**2 gene—such as *Zygosaccharomyces rouxii, K. lactis,* and *Ashbya gossypii*— supported levels of α2 repression comparable to the *S. cerevisiae* site (Figure 2A). *STE2 cis*-regulatory sequences taken from the *Candida* clade (*C. albicans* and *Pichia membranifaciens*) and the out-group species *Yarrowia lipolytica* failed to support repression in this assay (Figure 2A), consistent with the inference that in *C. albicans* and the *C. albicans*-*S. cerevisiae* ancestor, α2 does not repress the **a**sgs (Tsong, Tuch et al. 2006).

Full-length α2 ORFs from 8 species were fused to the *S. cerevisiae* α2 promoter and integrated into the genome in single copy (Figure 2B). α2 orthologs from species within the *Kluyveromyces* group repressed the **a**sg reporter comparable to levels observed for the *S. cerevisiae* protein (Figure 2B). In addition, the α2 ortholog of a species (*Z. rouxii*) that branches within the *Saccharomyces* group, but prior to the loss of **a**2, (Figure 1A) efficiently repressed the **a**sg reporter (Figure 2B). In contrast, α2 orthologs from *Candida* clade species failed to repress the reporter. The *C. albicans* α2 protein also failed to repress the *C. albicans* **a**sg *cis*-regulatory sequence (Figure 2C). These results show that changes in both the **a**sg *cis*-regulatory sequences and the α2 protein were both necessary for the switch in regulation and that the gain of α2 repression of the **a**sgs clearly preceded the loss of the **a**2 gene.

The clear trend from these experiments is that **a**sg *cis*-regulatory sequences and α2 proteins from the *Saccharomyces* and *Kluyveromyces* clades (Figure 1A) are competent to bring about repression, whereas those outside these clades are not. However, there is an important exception to this observed pattern. The *K. lactis* α2 protein failed to repress in this assay even though its *STE2 cis*-regulatory sequence is competent to bring about repression in this same assay (Figure 2B). To rule out the trivial possibility that α2 was misfolded or poorly expressed, we carried out a series of control experiments (Figure S1A). We will return to this unique feature of *K. lactis* later in this paper.

The evolution of a new function for α2

To investigate the molecular events that gave rise to α2 repression of the **a**sgs,we considered first the contribution of *trans* changes (coding sequence mutations in α2 or Mcm1).To identify regions of the α2 protein that may have been critical for the gain of α2-mediated repression, we quantified the levels of conservation across the α2 protein (Figure 3B). The α2 protein sequences from the hemiascomycete yeasts were divided into two groups: those that diverged prior to and those that diverged after the gain of α2 repression of the **a**sgs. In Figure 3B, high scores indicate conservation of those residues in the species group, whereas low scores indicate unconserved regions. Regions where the scores for the two groups are dissimilar reflect positions within α2 that experienced different levels of purifying selection in these two groups.

Much of the α2 protein has similar levels of conservation between the clades. This includes the 60 amino acid homeodomain (which mediates the sequence specific DNA-binding) (Hall and Johnson, 1987) and the 15 amino acid region of α2 that interacts with **a**1 (Mak and Johnson, 1993). DNA-binding and the interaction with **a**1 are functions of α2 that are required in all the clades considered, and their high sequences conservation reflects their high functional conservation. The α2 conservation traces diverged at two regions within the α2 protein, regions 1 and 3 (Figure 3A-C). Both regions displayed high levels of conservation in the *Saccharomyces-Kluyveromyces* lineages and low levels in the *Candida* lineage, implicating these regions in the evolution of α2 repression of **a**sgs. In fact, both regions are critical for α2 repression of the **a**sgs in *S. cerevisiae*; region 1 is responsible for recruiting the general repressor Tup1 (Komachi, Redd et al. 1994), and region 3 forms the interaction with Mcm1 (Vershon and Johnson 1993; Tan and Richmond 1998). The importance of the evolution of the Mcm1 interaction region in α2 (region 3) to the evolution of **a**sg repression is consistent with previous work using structural homology modeling (Tsong, Tuch et al. 2006).

To test these predictions directly, we designed a series of genetic swaps between the *C. albicans* and *S. cerevisiae* α2 proteins. The *S. cerevisae* α2 protein can be divided into five functional and structural regions (Figure 3A). We individually replaced each of these five regions of *S. cerevisiae* α2 with the homologous region of the *C. albicans* α2 protein and integrated (in single copy) the fusion proteins driven by the *S. cerevisiae* α2 promoter (Figure 3D). The ability of the modified α2 protein to repress expression was monitored using a reporter with a *S. cerevisiae* **a**sg or haploid specific gene *cis*-regulatory site in the promoter.

As predicted by the bioinformatic analysis, replacement of *S. cerevisiae* region 1 (Tup1 interaction) or region 3 (Mcm1 interaction) by the equivalent *C. albicans* sequences eliminated **a**sg repression. Also, as predicted, the swap of region 3 eliminated **a**sg repression, but left intact the protein's capacity for repression of the haploid specific genes. In contrast, the α2 functional region 1 swap protein (Tup1 interaction) failed to repress either the **a**sg reporter or the haploid specific gene reporter (Figure 3D). Replacing either functional region 1 or 3 with aligning sequence from another species *(Pichia pastoris)* that diverged prior to the gain of α2 repression at the **a**sgs gave similar results (Figure S1B). These observations show that the gain of **a**sg repression required the creation of two new functional regions within α2—a region that interacts with Mcm1 and a region that interacts with Tup1. In contrast to these two regions, the rest of the *S. cerevisiae* α2 protein sequence could be swapped for the homologous sequence from *C. albicans* α2 without a substantial effect on **a**sg repression. (Figure 3D).

Are the acquisition of the Tup1 and Mcm1 interaction regions was sufficient for α2 to acquire the capability to repress the **a**sgs? We swapped these functional regions from *S. cerevisiae* α2 into the *C. albicans* α2 protein and measured the ability of these hybrids to repress an **a**sg reporter. Neither region alone "rescued" the *C. albicans* protein; however, swapping both regions into *C. albicans* α2 together conferred the ability to repress the **a**sg reporter onto the hybrid protein (Figure 3E). These results demonstrate that the failure of the *C. albicans* α2 protein to repress the **a**sg reporter in *S. cerevisiae* reflect the inability of the protein to productively interact with both Tup1 and Mcm1. Consistent with this conclusion, swapping both of these regions into another *Candida*-group α2 protein (this one from *P. pastoris*) also conferred the ability to repress the **a**sgs onto that hybrid protein (Figure S1C). In summary, while two regions of α2 (regions 4 & 5) have been functionally conserved over large evolutionary distances (Figure 3B & D), two other regions (regions 1 & 3) evolved more recently in the ancestor of the *Saccharomyces*/*Kluyveromyces* groups (Figure 3B-C). These two recent additions are sufficient for α2 to gain its new function. This analysis illustrates how the evolutionary history of the α2 protein gave rise to its modular structural organization.

We also determined whether changes in Mcm1—the binding partner of α2—contributed to the evolution of **a**sg repression. To do this, we relied on ancestral gene reconstruction, an approach proven useful for testing evolutionary predictions (Thornton 2004). The strategy depends on the accurate protein alignments of the ortholog group of interest, followed by the calculation of amino acid probabilities at each position within the ancestral protein using a species or gene tree as a guide (Figure S2B). We reconstructed the complete MADS-box domain of Mcm1, the domain of Mcm1 that binds α2 (Mead, Zhong et al. 1996; Tan and Richmond 1998). Given the strong conservation of the Mcm1 MADS-box domain, all amino acid positions could be reconstructed within this domain with high accuracy in each ancestral protein. We synthesized a series of ancestral Mcm1 proteins and replaced the endogenous *S. cerevisiae* Mcm1 with them. Ancestral Mcm1 proteins dating back to the divergence of *S. cerevisiae-C. albicans* supported repression at levels equivalent to the modern *S. cerevisiae* Mcm1 (Figure S2). Thus, the gain of a new interaction between α2 and Mcm1 did not require changes in Mcm1. Instead, it appears that the evolution of the new protein-protein interaction was one-sided, with all the changes occurring in a short module of α2.

Integration of a new regulator into an existing regulatory network

Although the evolution of new protein-protein interaction modules in α2 was critical for the rewiring of the **a**sg network, the *cis*-regulatory sequences of the **a**sgs also evolved to become efficiently recognized by the α2 protein (Figure 2A). The similarities and differences between the **a**2-regulated (ancestral) and α2-regulated (derived) **a**sg *cis*-regulatory sequences have been described (Tsong, Tuch et al. 2006). The most striking similarities are the presence of a binding site for Mcm1 and the close relationship between the *cis*-regulatory sequences recognized by **a**2 and α2. Despite belonging to different transcription regulator superfamilies (HMG domain for **a**2 versus homeodomain for α2), both proteins recognize a core TGT sequence, with the outer nucleotides differing in their respective binding sites (Figure 3G). A major difference between the two regulatory sequences is in their symmetries. The *C. albicans* **a**2-regulated **a**sg binding sequence contains information specifying **a**2 binding on only one side of Mcm1. The *S. cerevisiae* α2 binding sequence, however, contains information on both sides of the Mcm1 binding site, specifying the binding of an α2 monomer on either side (Johnson and Herskowitz 1985).

In our next set of experiments, we examined in more detail the differences between the **a**2 and α2 recognition sequence and how the ancestral **a**2 site evolved to be recognized by α2. We found that *S. cerevisiae* α2 could repress *Kluyveromyces* group species **a**sg *cis*-regulatory sequences even though they varied significantly from the *S. cerevisiae* sites (Figure 3F). In fact, α2 efficiently repressed **a**sg *cis*-regulatory sequences (such as *Z. rouxii STE6* and *K. lactis STE2*) that contained precise **a**2 binding sites, as assessed by the Position Specific Scoring Matrix for **a**2 in the *Candida* clade (Figure 3G). In contrast, each **a**sg *cis*-regulatory sequence from a *Candida* group species failed to be repressed by *S. cerevisiae* α2 (Figure 3F), even when α2 was overexpressed (Figure S3). Thus, the ancestral **a**sg *cis*-regulatory sequences (recognized by **a**2) must have been converted to sites recognized by α2 along the *Saccharomyces-Kluyveromyces* lineage. To determine the minimum number of mutations necessary to convert an **a**2 site to a functional α2 site, we mutated three positions (positions 6, 26 and 27), from the *C. albicans RAM2 cis*-regulatory site, to their counterpart in the *S. cerevisiae* consensus sequence. Mutation of two of these nucleotides generated a construct that could be repressed by *S. cerevisiae* α2 (Figure 3H). Neither of these positions is highly constrained within the *Candida* group (Figure 3F-G). This conversion could occur without compromising the ancestral, positive regulatory mode because both proteins recognize the same core sequence (TGT). Specific bases to the "left" of the core are required for efficient **a**2 binding while specific bases to the "right" are required by α2 (Figure 3F). From these experiments we conclude that (1) *Candida* clade **a**-specific cis-regulatory sequences are recognized efficiently by **a**2, but not α2, (2) a small number of mutations (≤ 2) can convert an **a**2 site to an α2 site, and (3) these mutations occurred at positions that were likely under weak constraint in the ancestor.

The contribution of non-specific protein interactions to early intermediates

It is simple to envision how a couple of mutations could "convert" a single ancestral **a**sg *cis*-regulatory sequence into a sequence that can be recognized by α2. However, there are at least 7 **a**sgs in each species. And, as we discussed above, targeting of α2 to **a**sg *cis*-regulatory sequences also required the evolution of a new protein-protein interaction with Mcm1. How, then, did all of the gains required for this novel regulatory scheme arise? Did the Mcm1-α2 interaction evolve before or after the *cis*-regulatory changes? Or, did these events occur in concert?

To explore these questions, we mimicked two possible and extreme intermediate states in this evolutionary transition: the presence of the α2-Mcm1 protein-protein interaction without the *cis*-regulatory changes and the *cis*-regulatory changes without the α2-Mcm1 interaction. To create the first state, we replaced the *S. cerevisiae* **a**sg reporter with an **a**sg *cis*-regulatory sequence from the *Candida* clade (*C. albicans RAM2*). For the second state, we compromised the region of the *S. cerevisiae* α2 protein that binds Mcm1 by substituting it with the aligning sequence in the *C. albicans* protein. When the *C. albicans RAM2 cis*-regulatory sequence was tested with wild-type *S. cerevisiae* α2, we did not observe repression, even when α2 was over-expressed. However, when the Mcm1 interaction region was disrupted but the *S. cerevisiae cis*-regulatory sequence was used, we did observe repression when α2 was overexpressed. (Figure 4A)

We next determined how the α2 protein lacking the Mcm1 interaction region could still repress an **a**sg reporter, albeit weakly. In principle, either the "ancestral" α2 could bind the **a**sg reporter independently of Mcm1 or Mcm1 could stabilize ancestral α2 binding through non-specific protein-protein interactions. To distinguish between the models, we tested for repression of an **a**-specific *cis*-regulatory sequence in which the Mcm1 *cis*-regulatory site was destroyed by mutation (Figure 4B). (Mcm1, an essential protein, cannot be deleted from the cell.) Using this reporter, overexpression of a modified α2 protein that lacks the Mcm1 interaction region failed to show any detectable repression (Figure 4B). Thus, it appears that the second model best accounts for our results: even before the evolution of a specific Mcm1-interaction region, binding of the "ancestral" α2 was stabilized by its proximity to Mcm1. These results suggest a model where the effects of fortuitous *cis*-mutations, which stabilized α2 binding to DNA, would have been amplified by the contribution of non-specific interactions with Mcm1 during the earliest steps in the evolution of α2 repression at the **a**sgs.

We hypothesize that once a more optimized Mcm1-α2 protein interaction formed, α2 could have occupied *cis*-regulatory sequences that deviate from its preferred sequences. These types of sites may have occurred in intermediates and we modeled such an intermediate by mutating a single, key base pair in the *S. cerevisiae STE2 cis*-regulatory sequence. Even with a mutated α2 binding site, we find that when α2 is overexpressed, it can mediate repression, but only if the Mcm1 interaction region of α2 is present (Figure 4C). Thus, a protein-protein interaction with Mcm1 can stabilize the binding of α2 to imperfect *cis*-regulatory sequences; such sequences may have been present in early, evolutionary intermediates.

If these ideas are correct, then the changes in *cis*-regulatory sequences and the evolution of this new protein-protein interaction are linked and must have evolved together (see (Tuch, Li et al. 2008; Wagner 2008). An attractive feature of this co-evolution model is that the interaction energy needed for the α2 and Mcm1 proteins to occupy an **a**sg *cis*-regulatory sequence can be distributed between the protein-protein and protein-DNA interactions, enabling all the **a**sgs to come under weak influence by α2 and then tuned individually through changes in each gene's *cis*-regulatory sequence.

*Hybrid regulation of **a**sgs by both **a**2 and α2 occurs in modern species*

The experiments described here and by Tsong et al., 2006 indicate that the control of **a**sg expression passed through a hybrid regulatory state in which positive control by **a**2 and negative control α2 operated together. One can envision two, non-mutually exclusive types of such hybrid regulation. In the first, a given **a**sg would be both repressed by α2 in α cells and activated by **a**2 in **a** cells. In the second, regulation would be at the network level; some **a**sgs would be activated by **a**2 in **a** cells and other **a**sgs would be repressed by α2 in α cells. Both types of hybrid regulation would ensure that each **a**sg is expressed only in **a** cells. We next investigated the possibility that some form of hybrid regulation still exists in modern species. We chose to examine *L. kluyveri* and *K. wickerhamii* because both have an intact **a**2 gene (Butler, Kenny et al. 2004), and the α2 protein of both species is able to repress a *S. cerevisiae* **a**sg *cis*-regulatory site (Figure 1A and 2B).

In *L. kluyveri,* a genome-wide ChIP of **a**2 was performed in **a** cells (Figure 5A, C, E and S4). Ten peaks of **a**2 binding met our enrichment cut-offs, and six of these peaks were upstream of genes whose orthologs are **a**sgs in either *C. albicans* or *S. cerevisiae* (*AGA2*, *ASG7*, *AXL1*, *BAR1*, *STE2*, and *STE6*) (Tsong, Miller et al. 2003; Galgoczy, Cassidy-Stone et al. 2004). To determine if these genes and the genes associated with the remaining four peaks are expressed in an **a**-specific pattern, RT-qPCR was performed using wild-type **a** cells and wild-type α cells (Figure S5A). We also tested the gene *RAM1* because *RAM1* is an **a**sg in *C. albicans* (Tsong, Miller et al. 2003), and its peak of **a**2 binding fell just below our significance threshold. Using this data, we defined the following nine genes as *L. kluyveri* **a**sgs: *AGA1, AGA2, ASG7, AXL1, BAR1, RAM1, STE2, STE6,* and *STE14*. Two of these genes, *STE14* and *AGA1* are **a**sgs in *L. kluyveri* but not in either *S. cerevisiae* or *C. albicans*; the others are **a**sgs in at least two of the three species. (Three genes associated with **a**2 binding in *L. kluyveri* (*ELA1*, *TID3*, and *SAKL0E14784g*) did not show **a**sg expression under any condition we tested and were excluded from further tests.) Transcript levels of all nine *L. kluyveri* **a**sgs were decreased when **a**2 was deleted (Δ*MAT**a**2*), indicating that **a**2 activates these genes by binding to their *cis*-regulatory sequences (Figure 5G).

Next, full genome ChIP of myc-tagged α2 in α cells was used to ascertain its role, if any, in the regulation of **a**sgs, in *L. kluyveri* (Figure 5B, D, F and Figure S4). In α cells, binding peaks were observed upstream of two genes—the **a**sgs *AGA1* and *AGA2* (Figure 5B and D). These peaks are centered over the same region of DNA as the **a**2 binding peaks observed in **a** cells, showing that the two regulators associate with the same region of DNA but in different cell types. This result is consistent with the analysis described above showing that the two regulators have overlapping DNA binding specificities and each forms a protein interaction with Mcm1 (Figure 3G). To test whether *AGA1* and *AGA2* are repressed by α2, we performed RT-qPCR in wild-type α cells and in α2-deletion α cells (Δ*MATα2*) (Figure 5H). The transcript abundance of both of these genes increased indicating that α2 represses these genes in α cells. The remaining seven **a**sgs were also tested by RT-qPCR and determined not to be targets of α2 repression in these conditions (Figure 5H). Taken together, these results indicate that all nine of the *L. kluyveri* **a**sgs are targets of direct **a**2 activation in **a** cells and that two of them are also targets of direct α2 repression in α cells. Thus, in *L. kluyveri,* two of the **a**sgs are regulated in a hybrid fashion. The results also show that, for these two genes, **a**2 and α2 act through association with the same DNA sequence in the two cell types.

The other species chosen for this analysis, *K. wickerhamii*, is described in Figure S6. The results indicate that at least two **a**sgs are regulated in a hybrid fashion in *K. wickerhamii*. We note that the genes that are hybrid-regulated in *K. wickerhamii* are not the same genes that are hybrid-regulated in *L. kluyveri* (summarized in Figure 7C).

Gains and losses in the **a***sg network*

In addition to changes in the overall form of regulation, we find that the **a**sg network has gained and lost individual target genes over the hemiascomycete lineage. We believe this can be accounted for by the formation and destruction of *cis*-regulatory sequences. For instance, we found that *STE14* is an **a**sg in *L. kluyveri* but not in the other species examined and that *AXL1* is an **a**sg in many species but not *S. cerevisiae* (Figure 7C, Table S2 and S3 and (Tsong, Miller et al. 2003; Galgoczy, Cassidy-Stone et al. 2004; Booth, Tuch et al. 2010)).

K. lactis α2 lost the ability to repress **a**sgs

The dairy yeast *K. lactis* diverged from *S. cerevisiae* after the gain of **a**sg repression, and it retains many of the *cis* and *trans* characteristics indicative of a hybrid form of regulation where both **a**2 with α2 are active (Tsong, Tuch et al. 2006). Yet, as noted above, the *K. lactis* α2 protein is unable to repress the **a**sgs when moved into *S. cerevisiae* (Figure 2B-C).

To determine whether α2 represses the **a**sgs in *K. lactis* itself, we utilized gene expression profiling to compare transcript levels of wild-type **a** and wild-type α cells to Δ**a**2 **a** cells and Δα2 α cells, respectively. Deletion of α2 in α cells did not have an effect on transcript levels of any of the *K. lactis* **a**sgs (Figure 6E and Figure S5B) nor did it affect the expression of other genes in *K. lactis* (data not shown). We confirmed this result by measuring transcript levels of **a**sgs by RT-qPCR (data not shown). In contrast, deleting **a**2 in **a**-cells resulted in decreased expression of nearly all of the *K. lactis* **a**sgs (Figure 6E). Consistent with these results, **a**2 was found to be bound upstream of the *K. lactis* **a**sgs (Figure 6A, C and data not shown) but α2 binding was not detected at the **a**sgs or any other gene in α cells (Figure 6B, D and data not shown). (As a control, *K. lactis* α2 binding is observed at the haploid specific genes when α2 and **a**1 are expressed together (Booth, Tuch et al. 2010).) Thus, although *K. lactis* has many of the hallmarks of hybrid regulation (in particular, its **a**sg *cis*-regulatory sequences support repression by *S. cerevisiae* (Figure 2A), α2 does not repress the **a**sgs in this species.

Comparison of the α2 sequences from multiple species pointed to a likely cause of the inability of the *K. lactis* α2 to repress the **a**sgs: amino acid residue 136 in *K. lactis* is an asparagine, but in all repressing-competent α2 proteins it is a small, hydrophobic residue, either a valine or leucine (Figure 3C). This position has been shown to be important for the interaction between α2 and Mcm1 (Mead, Zhong et al. 1996; Tan and Richmond 1998). Using the *S. cerevisiae* reporter assay, we tested this idea explicitly and found that mutating this single residue in the *K. lactis* α2 protein to a valine (N136V) restored its function as a repressor (Figure 6G). The simplest interpretation of these observations is that the *K. lactis* α2 protein recently acquired a mutation that compromised its ability to interact with Mcm1 thereby destroying the derived (repression) mode of **a**sg regulation and reverting to the ancestral (positive) mode. The evolutionary path by which this amino acid substitution likely occurred is explored in detail in Figure S7.

Discussion

The regulation of a set of cell-type specific genes, the **a**sgs, has changed over evolutionary time in the hemiascomycete branch of the fungal lineage. Based on data from numerous approaches, we describe the likely evolutionary path for the change in the mechanism by which the **a**sgs are regulated. We provide strong experimental evidence for an intermediate hybrid regulatory state in which **a**2 and α2 both participated in the cell-type regulation of the **a**sgs, and we show that this hybrid state resolved in several distinct ways along the lineages to modern species, generating a diversity of network structures (summarized in Figure 7A).

The gain of α2 repression at the **a**sgs required that α2 navigate a constrained regulatory landscape. As a result, this evolutionary path exploited multiple features of the existing network that both stabilized early intermediates and limited the number of mutations required to evolve this new function. We also show that protein modularity minimized the pleiotropy of the evolved features of the new regulatory mode. This work provides both a mechanistic account of how a particular transcription regulator evolved a new function and insights into the molecular origins of the extraordinary flexibility of transcriptional regulatory network architectures that appear across modern species.

In this discussion we first outline the key features of the ancestral network that were exploited (that is, exaptations) in the evolution of α2-repression of the **a**sgs. We next discuss the concerted changes in the *cis*-regulatory sequences and the *trans* regulators that enabled formation of the new mode of regulation. Third, we consider the consequences of the intermediate hybrid regulatory state and its role in the network diversity observed in modern species. Finally, we discuss the relative importance of adaptation and neutral drift to the diversification of gene regulatory networks.

Exploitation of ancestral network components

Several key features of the derived form of regulation (repression of the **a**sgs) were in place prior to its evolution. For instance, the new mode of regulation requires that the repressor be expressed in α and **a**/α cells, but not in **a** cells. For α2, this is true for virtually every species in the hemiascomycetes and reflects its deeply conserved function: it forms a heterodimer with **a**1 to regulate the haploid specific genes in **a**/α cells (Strathern, Hicks et al. 1981; Tsong, Miller et al. 2003; Booth, Tuch et al. 2010). Thus, the expression pattern necessary for α2 to act as a repressor of the **a**sgs was already present in the ancestor.

In contrast to the popular model wherby new *cis*-regulatory sequences arise *de novo* in unused regions of promoters, α2 exploited features of the existing **a**sg *cis*-regulatory sequences (Tsong, Tuch et al. 2006). The monomers of **a**2 and α2 have related DNA-binding specificities (Figure 3G) despite belonging to different transcription regulator families (HMG box vs. homeodomain, respectively). This intrinsic overlap in DNA-binding specificities minimized the number of *cis*-regulatory mutations required for the transition: only two point mutations are required to convert an optimal **a**2 recognition sequence to an optimal α2 recognition sequence (Figure 3H). Moreover, we have shown that sequences exist in modern species that are efficiently recognized by both proteins (Figures 5, S4 and S6), thus further reducing the potential fitness barriers to this transition.

In addition to the exploitation of **a**2 *cis*-sequences, the binding of α2 to the ancestral sequences was stabilized by the presence of a neighboring DNA-bound protein, Mcm1. We provide evidence for a model where the ancestral presence of Mcm1 at the *cis*-regulatory sites of the **a**sgs stabilized α2 DNA binding in early evolutionary intermediates through weak, relatively non-specific protein-protein contacts (Figure 4A and B). Subsequently, the protein-protein interaction became stronger and more specific through changes in the α2 protein, which stabilized the binding of Mcm1 and α2 to each other and to DNA. We have shown that the evolution of this specific interaction between Mcm1 and α2 was asymmetric: the α2 protein underwent numerous changes in a previously unconstrained region allowing it to recognize an existing surface of the ancestral Mcm1; therefore, no changes were necessary in Mcm1 (Figure 3B-E). Thus, from the earliest steps in this evolutionary transition, the interaction energy necessary to stabilize α2 binding was shared out between protein-protein and protein-DNA contacts. The exploitation of ancestral *cis* and *trans* features strongly guided the evolutionary trajectory of α2 (through stabilizing early intermediates) by minimizing the number of changes necessary.

Constraint and the evolution of novelty by cis *and* trans *changes*

Although several key network features needed for the evolution of α2-repression of the **a**sgs were already present in the ancestor, changes in both the *cis*-regulatory sequences and the α2 protein needed to occur for efficient **a**sg repression. The gain and loss of *cis*-regulatory sequences are readily acknowledged as major contributors to evolutionary novelty, but changes in the transcription regulators themselves are often described as less prevalent, particularly in the absence of gene duplication (Carroll 2005; Wray 2007). For example, it is frequently said that changes in transcription regulators will tend to be rare because they are pleiotropic—affecting the regulation of many genes simultaneously and likely disrupting existing networks.

The gain of function of α2 described here occurred within the context of a pre-existing, deeply conserved regulatory landscape: the regulation of the haploid specific genes by the **a**1-α2 heterodimer (Herskowitz 1989; Hull and Johnson 1999; Booth, Tuch et al. 2010). The modularity of the α2 protein made it possible to gain a new function (repression of the **a**sgs) without compromising its ancestral function (repression of the haploid specific genes). Indeed, it seems likely that the only permissible evolutionary trajectories for the α2 protein to gain a new function would require that its ancestral function be preserved. How did this occur?

Two regions of the α2 protein—the DNA-binding homeodomain and the **a**1 interaction region—are needed for its ancestral function and are preserved, in sequence and function, through stabilizing selection across the entire hemiascomycete lineage (Figure 3B & D). The protein modules that more recently evolved to make **a**sg repression possible (regions 1 and 3, Figure 3B, C, and E) are short (~10) stretches of amino acids that developed within unconstrained regions of the ancestral protein (Figure 3B and C). The evolution of short, linear protein interaction regions spatially isolated from the ancestral functions bypassed the potential pleiotropic constraints on regulator evolution. We note that the gain of new functional modules in unused portions of the ancestral protein is akin to the acquisition of new *cis*-regulatory sequences at unconstrained positions in non-coding sequence. More generally, the modular structure of modern transcription regulators is likely the result of the sequential addition of new functions in previously unconstrained regions of the proteins, as described here.

Hybrid intermediates and the diversification of regulatory networks

As we have described, the path to the gain of α2-repression of the **a**sgs occurred while the ancestral form of **a**-specific regulation (activation by **a**2) was still extant (Tsong, Tuch et al. 2006). Thus, both forms of regulation existed together in the ancestor of the *Kluyveromyes* and *Saccharomyces* clades. We propose that this hybrid regulatory intermediate made possible the subsequent diversification of the **a**sg regulatory network architectures without a loss in regulation. Based on evidence from several modern species, we found that the hybrid regulatory state has diversified (resolved) in three directions:

- *Retention of both modes of regulation:* We showed that two modern species, *K. wickerhamii* and *L. kluyveri*, have retained both the ancestral (**a**2 activation) and derived (α2 repression) modes of regulation of the **a**sgs (Figures 5 & S6). Two additional species, *Z. rouxii* and *A. gossypii*, also possess α2 proteins that repress **a**sg expression (Figure 2B) and both appear to have functional **a**2 genes. Thus, we favor the hypothesis that these two species also retain some form of the hybrid regulatory state.
- *Loss of the ancestral mode of regulation: S. cerevisiae* and other post-whole genome duplication species regulate their **asg**s using the repressor α2 exclusively. Indeed, the gene coding for the activator **a**2 (the ancestral regulator) has been lost

from these species (Butler, Kenny et al. 2004); thus, the ancestral mode has been discarded.

- *Loss of the derived mode of regulation: K. lactis* appears to have lost α2 repression of the **a**sgs through a recent, single amino acid change in the α2 protein. The α2 protein of the nearby branching species *Kluyveromyces marxianus* also has a mutation at this same position (Figure 3C), although the substituted amino acid is different in the two species. In *K. lactis* (and presumably *K. marxianus*), the **a**sgs appear to be regulated by **a**2 alone, with the derived mode no longer in use.

We suggest hybrid regulatory states, such as the state described here, represent 'high potential states' for evolutionary change as they have the ability to resolve in several directions without destroying the overall logic of regulation (Figure 7B). Akin to gene duplication, the formation of a hybrid regulatory state generates a partially redundant intermediate that allows for diversification without a loss of the original function or regulatory logic (Tanay, Regev et al. 2005). Within the hybrid regulatory state, network reversion remains a permissible evolutionary trajectory. The reversion to an ancestral regulatory mode that we have described in *K. lactis* is not a strict molecular reversal. Instead, the *K. lactis* α2 protein acquired a mutation that inactivates the derived function while maintaining its ancestral function, haploid specific gene repression as a heterodimer with **a**1.

Our results also show that, over the evolutionary time period considered in this paper, a subset of **a**sgs moved in and out of the network through the gains and losses of *cis*-regulatory sequences (summarized in Figure 7C). Although some genes are expressed **a**-specifically in all species (e.g. those encoding pheromones and pheromone receptors), others are not. This implies that for the **a**sgs to undergo a transition from one regulatory mode to another, not all genes within the network would need to experience this switch in regulation. The looser requirements for the regulation of some genes in a network may facilitate changes in the mode of regulation of a network, as not all genes would have to be carried along during the initial phases of the switch.

Adaptive and neutral forces in regulatory evolution

Selection can only act on the output of a transcription regulatory network; if an evolutionary path exists between different regulatory architectures with near-identical spatial pattern, dynamic range, and kinetics of expression, then the network can be predicted to drift between these different solutions over evolutionary time (Lynch 2007). The hybrid state we have described spawned a range of evolutionary outcomes (activation, repression or hybrid), each with different regulatory circuit architectures. In all cases, however, the overall logic of regulation (**a**sgs ON in **a** cells and OFF in the other two cell types) has been preserved. It is possible that each of the different forms of regulation we observed produce different dynamic ranges or kinetics of expression and that these qualities have been selected for on a gene-by-gene basis as different yeast species diversified. However, we favor the simpler model where the regulatory diversification following the formation of the hybrid regulatory state occurred largely through neutral, non-adaptive, drift. In other words, the network could drift between states where the dynamic range of regulation generally remained the same but the relative contributions of the ancestral and derived modes differed through the strengthening and weakening of protein-protein and protein-DNA interactions. The range of network structures observed in modern species would simply reflect the "breathing" of the hybrid regulatory network.

In contrast to the neutral model we favor for network diversification from the hybrid state, we currently favor the idea that the formation of the hybrid state was itself adaptive. For one thing, the gain of **a**sg repression to form the hybrid state required a reasonably large number of mutational events, both in *cis* and *trans*. For instance, the gain of two new protein interaction modules within α2 (one for Tup1 and one for Mcm1) involved greater than two-dozen amino acid changes and it seems unlikely that such a large number of amino acid changes that produce a new biochemical function could have reached fixation without directional selection. We cannot know for certain what adaptive value the invention of **a**sg repression had, if any, for the ancestor of the *Kluyveromyces* and *Saccharomyces* clades. However, the gain of repression at this gene set may have been a necessary regulatory response to another newly evolved trait in this ancestor, the gain of silent mating cassettes (Butler, Kenny et al. 2004). These additional mating cassettes—containing copies of the mating-type regulates—are silenced in *S. cerevisiae* by heterochromatin. The risk is that simultaneous expression of both sets of haploid mating-type genes can lead to cell cycle arrest. Thus, leaky silencing of the mating activator **a**2 in the wrong cell-type may have provided a strong selective pressure for the gain of the repression mode of **a**sg control. Together, these arguments are not conclusive, but they are consistent with the idea that positive selection played a role in the gain of α2 repression of the **a**sgs and the formation of the hybrid intermediate, and that the successive circuit diversification was non-adaptive.

Irrespective of the potential role of selection, a hybrid regulatory state can be short-lived (as in the ancestor of *S. cerevisiae*) or exceedingly long-lived (as in *L. kluyveri* and *K. wickerhamii*). We propose that the creation of hybrid regulatory states serves as a general model to rationalize the many examples of network-wide transcriptional regulatory divergence that have been observed among species.

Experimental Procedures:

Identification of Gene Orthologs and Upstream Regulatory Sequences

Orthologs of experimentally identified **a**sgs (Galgoczy, Cassidy-Stone et al. 2004) (Tsong, Miller et al. 2003) were identified and confirmed using BLAST. To identify a Position Specific Scoring Matrix (PSSM) for α2-repression (derived), we submitted to MEME the 600 base pairs upstream of the **a**sgs from *S. cerevisiae*, *Saccharomyces mikatae, Saccharomyces paradoxus,* and *Saccharomyces bayanus.* Similarly, sequences from *C. albicans, Candida dubliniensis,* and *Candida tropicalis* were used to calculate a PSSM for **a**2-activation (ancestral). The 600 base pairs upstream of each **a**sg were scanned to identify the **a**sg *cis*-regulatory sequences of all genome sequenced hemiascomycetes using MAST (Bailey, Boden et al. 2009). See Extended Experimental Procedures for details.

Strain Construction

A complete list of all strains used in this study can be found in Table S5. The primers used to generate and confirm these strains are listed in Table S6. For details regarding strain and plasmid construction see Extended Experimental Procedures.

β-galactosidase Assays

β-galactosidase assays were performed using a standard protocol (Guarente and Ptashne 1981). Strains were grown in selective media to maintain transformed plasmids. For each strain, colonies were grown overnight, diluted, and allowed to reach late log phase. Cells were harvested and permeabilized, and activation assays were performed.

Quantification of Conservation Scores within α2

α2 orthologs were aligned using MUSCLE (Edgar 2004). The genetic diversity spanned by the *Saccharomyces*-*Kluyveromyces* and *Candida* clade is similar (Taylor and Berbee 2006), however, we removed from our analysis a subset of closely related sequences from the *Saccharomyces-Kluyveromyces* species to normalize the levels of conservation between the two groups. The displayed amino-acid conservation was calculated using the PAM250 amino-acid substitution matrix (Henikoff and Henikoff 1992). The displayed curve (Figure 3B) has been smoothed by averaging each conservation score with the scores of adjacent residues. See Extended Experimental Procedures for details.

RNA Isolation and cDNA Preparation

RNA was isolated from yeast cultures using hot phenol/chloroform extraction. cDNA was prepared using SuperScript II (Invitrogen). Additional details can be found in the Extended Experimental Procedures.

Gene Expression Arrays

K. lactis cDNA was hybridized to a custom Agilent array. All data has been deposited in NCBI GEO at accession number (GSE39027). cDNA labeling, hybridization and data analysis are described in the Extended Experimental Procedures.

Chromatin Immunoprecipitation

C-terminally myc tagged **a**2 and α2 proteins were created for ChIP. Tagged (experimental) and untagged (control) strains were grown, harvested and lysed. Chromatin was precipitated with commercially available anti-myc or anti-HA antibodies. The DNA was amplified, labeled and competitively hybridized to custom Agilent tiling oligonucleotide arrays. Display, analysis and identification of binding events were performed with MochiView (Homann and Johnson 2010). Details are found in the

Extended Experimental Procedures. Data has been deposited in NCBI GEO at accession numbers GSE38919 for *K. lactis* and (GSE39007) for *L. kluyveri*.

Quantitative PCR

A complete list of all primers used for qPCR is found in Table S6.

Acknowledgements

We are grateful to Chiraj Dalal, Clarissa Nobile, and Brian Tuch for critically reading the manuscript and Jennifer Garcia, Victor Hanson-Smith, Oliver Homann, Quinn Mitrovich, and Brian Tuch for their generous and helpful advice concerning experimental and computational procedures. We thank Peter Philippsen for kindly sharing genomic DNA and sequences from *Ashbya gossypii*. This work was supported by RO1 GM057049 from the National Institute of Health. The authors declare no conflict of interest.

C.R.B. performed reporter assays and the Mcm1 ancestral gene reconstruction. L.N.B. performed experiments in *L. kluyveri* and *K. wickerhamii*. L.N.B. and T.R.S. performed *K. lactis* experiments. Data was analyzed and computation experiments performed by C.R.B., L.N.B., and T.R.S. All authors contributed to the design of the study and wrote the paper.

Figure 1

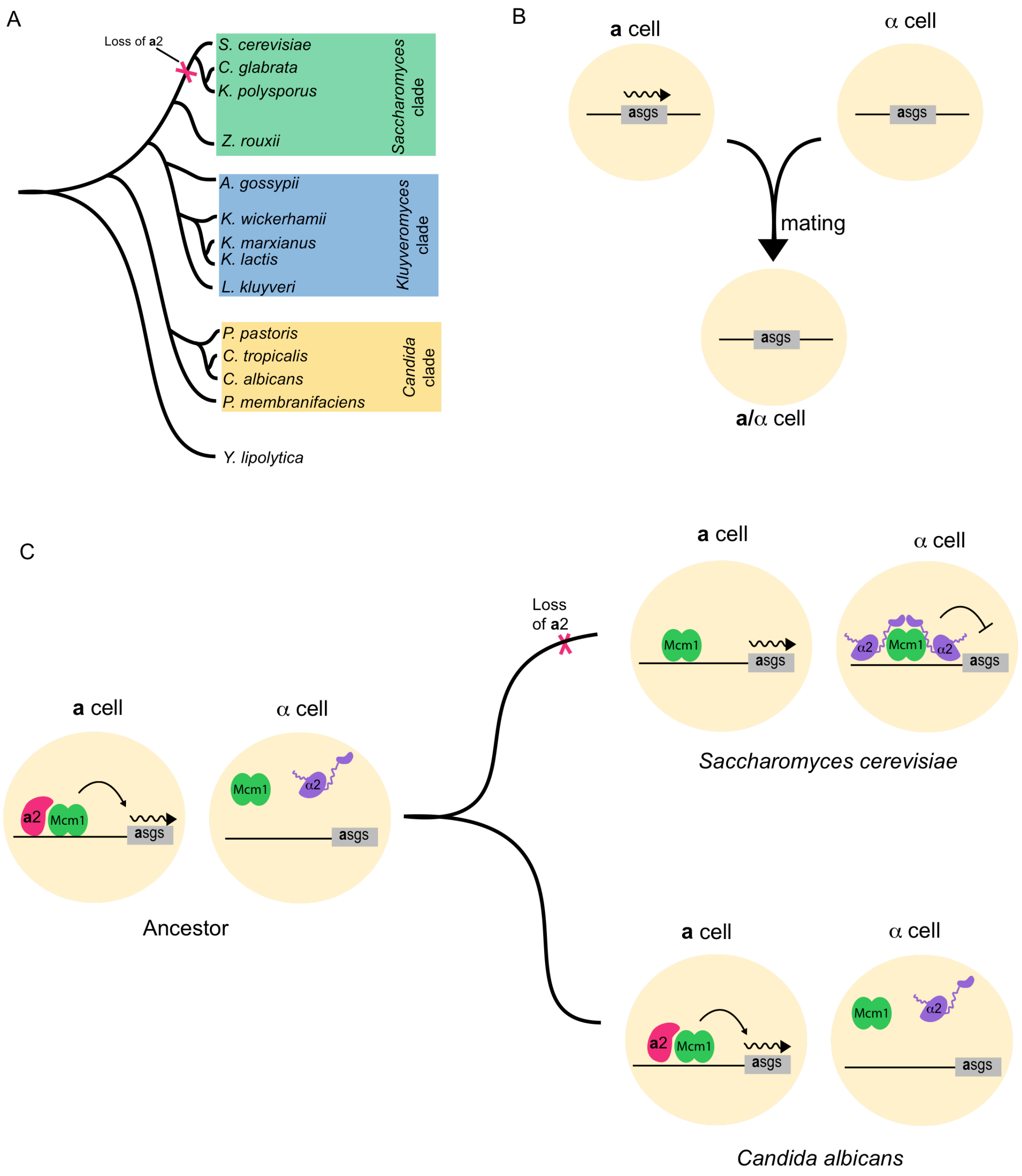


Figure 1: Cell-type specification in the hemiascomycetes

(A) Three hemiascomycete clades are considered—*Candida*, *Kluyveromyces* and *Saccharomyces*. The *Saccharmoyces* clade includes the pre-whole genome duplication species *Zygosaccharomyces rouxii* and the post-whole genome duplication species that lack an **a**2 gene (loss event indicated by a pink X). (B) The hemiascomycete yeasts have three cell types; the mating competent **a** and α cells and the product of their mating, an **a**/α cell. **a** cells express a set of genes called the **a**sgs (**a**sgs) (Herskowitz 1989). (C) In *C. albicans* and the ancestor, the **a**sgs are activated by Mcm1 (present in all cell types) and **a**2 (present only in **a**-cells) (Tsong, Miller et al. 2003). In *S. cerevisiae*, the **a**sgs are specified using Mcm1 a cell-type specific repressor, α2 (Johnson and Herskowitz 1985; Keleher, Goutte et al. 1988).

Figure 2

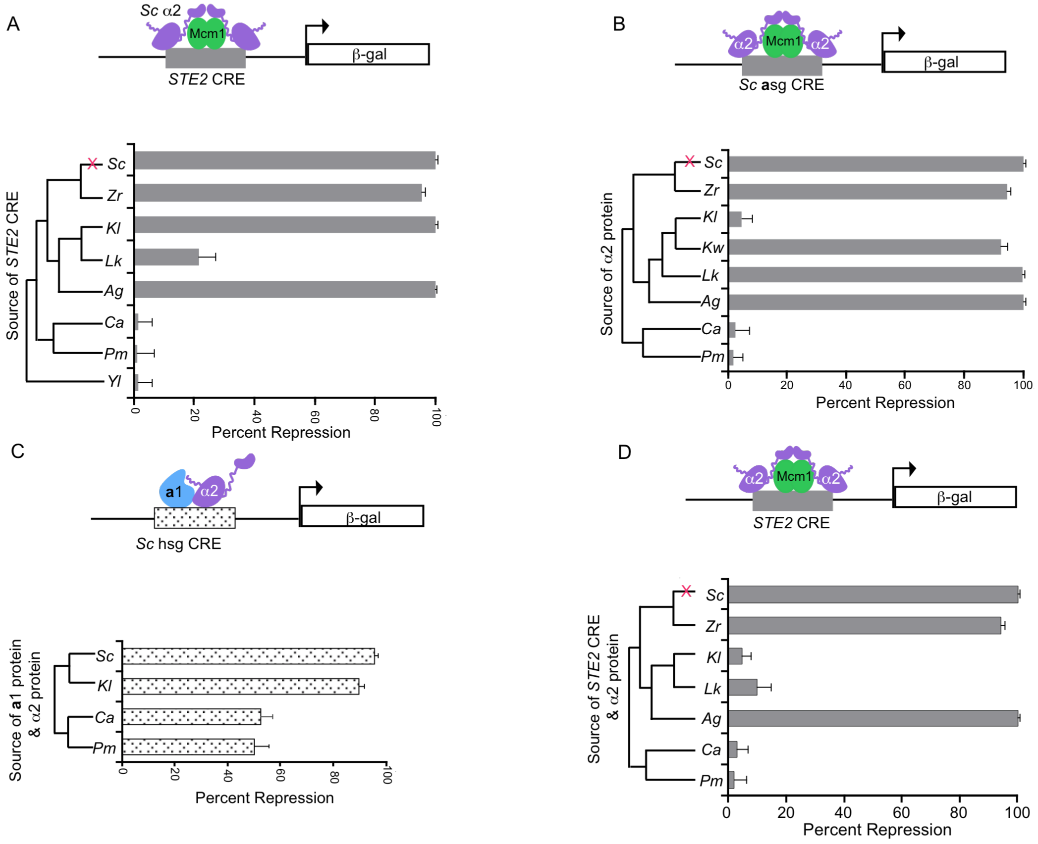


Figure 2: α2 repression of the asgs evolved prior to the divergence of *Saccharomyces* and *Kluyveromyces*

(A) The **a**sg *cis*-regulatory sequence of the α-pheromone receptor gene *STE2* from *S. cerevisiae* (*Sc)* and species that branch prior to the loss of the **a**2 gene, *Z. rouxii (Zr), K. lactis (Kl), L. kluyveri (Lk), A. gossypii (Ag), C. albicans (Ca), P. membranificians (Pm),* and *Y. lipolytica (Yl)* were inserted into a reporter construct to assay repression. Percent repression was determined by transforming constructs into *S. cerevisiae* **a**-cells (no α2) and α-cells (α2 present). (B) α2 protein coding sequence from a variety of hemiascomycete species including *K. wickerhamii (Kw)* were fused to the endogenous *S. cerevisiae* α2 promoter and integrated into the genome of a *S. cerevisiae MATΔ* strain. "Trans-species" α2 proteins were then assayed for their ability to repress the *S. cerevisiae STE2* **a**sg reporter. (C) Trans-species α2 proteins were combined with the *STE2 cis*-regulatory sequence reporter constructs from the same species and assayed for repression in a *MAT*Δ background. All values reported are a mean (*n*=3) and standard error of the mean.

Figure 3

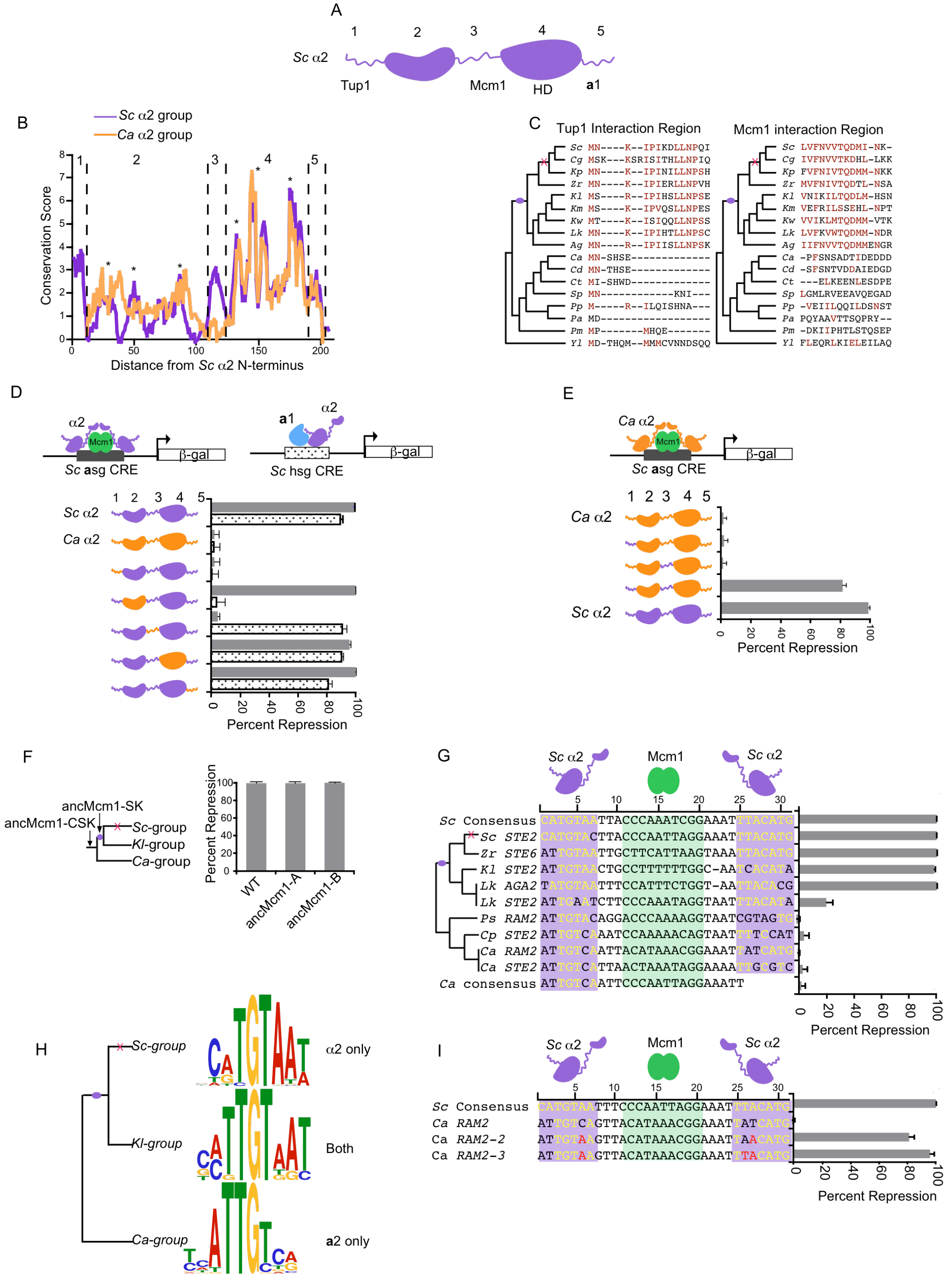

A
1
2
3
4
5
Sc α2
Tup1
Mcm1
HD
a1
B
Sc α2 group
Ca α2 group
Conservation Score
Distance from Sc α2 N-terminus
0
1
2
3
4
5
6
7
8
0
50
100
150
200
*
C
Tup1 Interaction Region
Sc MN----K--IPIKDLLNPQI
Cg MSK---KSRISITHLLNPIQ
Kp MN----K--IPINILLNPSH
Zr MN----K--IPIERLLNPVH
Kl MS----R--IPIHSLLNPSE
Km MS----K--IPVQSLLNPES
Kw MT----K--ISIQQLLNPSS
Lk MN----K--IPIHTLLNPSC
Ag MN----R--IPIISLLNPSK
Ca MN-SHSE-------------
Cd MN-THSE-------------
Ct MI-SHWD-------------
Sp MN------------KNI---
Pp M-----R--ILQISHNA---
Pa MD------------------
Pm MP-------MHQE-------
Yl MD-THQM--MMMCVNNDSQQ
Mcm1 interaction Region
Sc LVFNVVTQDMI-NK-
Cg IVFNVVTKDHL-LKK
Kp FVFNIVTQDMM-NKK
Zr MVFNIVTQDTL-NSA
Kl VNIKILTQDLM-HSN
Km VEFRILSSEHL-NPT
Kw VVIKLMTQDMM-VTK
Lk LVFKVWTQDMM-NDR
Ag IIFNVVTQDMMENGR
Ca -PFSNSADTIDEDDD
Cd -SFSNTVDDAIEDGD
Ct ---ELKEENLESDPE
Sp LGMLRVEEAVQEGAD
Pp -VEIILQQILDSNST
Pa PQYAAVTTSQPRY--
Pm -DKIIPHTLSTQSEP
Yl FLEQRLKIELEILAQ
D
α2
Mcm1
Sc asg CRE
β-gal
a1
α2
Sc hsg CRE
β-gal
1 2 3 4 5
Sc α2
Ca α2
0
20
40
60
80
100
Percent Repression
E
Ca α2
Mcm1
Sc asg CRE
β-gal
1 2 3 4 5
Ca α2
Sc α2
0
20
40
60
80
100
Percent Repression
F
ancMcm1-SK
ancMcm1-CSK
Sc-group
Kl-group
Ca-group
Percent Repression
0
20
40
60
80
100
WT
ancMcm1-A
ancMcm1-B
G
Sc α2
Mcm1
Sc α2
5
10
15
20
25
30
Sc Consensus CATGTAATTACCCAAATCGGAAATTTACATG
Sc STE2 CATGTACTTACCCAATTAGGAAATTTACATG
Zr STE6 ATTGTAATTGCTTCATTAAGTAAATTACATG
Kl STE2 ATTGTAACTGCCTTTTTTGGC-AATCACATA
Lk AGA2 TATGTAATTTCCATTTCTGGT-AATTACACG
Lk STE2 ATTGAATCTTCCCAAATAGGTAATTTACATA
Ps RAM2 ATTGTACAGGACCCAAAAGGTAATCGTAGTG
Cp STE2 ATTGTCAAATCCAAAAACAGTAATTTTCCAT
Ca RAM2 ATTGTCAATTACATAAACGGAAATTATCATG
Ca STE2 ATTGTCATTAACTAAATAGGAAAATTGCGTC
Ca consensus ATTGTCAATTCCCAATTAGGAAATT
0
20
40
60
80
100
Percent Repression
H
Sc-group
Kl-group
Ca-group
α2 only
Both
a2 only
I
Sc α2
Mcm1
Sc α2
5
10
15
20
25
30
Sc Consensus CATGTAATTTCCCAATTAGGAAATTTACATG
Ca RAM2 ATTGTCAGTTACATAAACGGAAATTATCATG
Ca RAM2-2 ATTGTAAGTTACATAAACGGAAATTAACATG
Ca RAM2-3 ATTGTAAGTTACATAAACGGAAATTTACATG
0
20
40
60
80
100
Percent Repression

Figure 3: The *cis* and *trans*-evolution underlying the gain of a new function for α2

(A) Structured regions of *S. cerevisiae* α2 are displayed as globular, whereas, unstructured regions are displayed as curved lines. (B) Conservation scores for the α2 protein across the *Saccharomyces-Kluyveromyces* group (*Sc)* or the *Candida*-group *(Ca)*. The vertical dashed lines correspond to the edges of the modular regions within the α2 protein. The positions of the three structurally predicted helices within regions 2 and 4 are marked (*) (C) The MUSCLE alignment for regions 1 and 3 are displayed. (D) *S. cerevisiae* α2 modules were swapped for the homologous regions from the *C. albicans* α2 protein. Each construct was genome-integrated in a *MATΔ* background and assayed for the ability to repress the *S. cerevisiae STE2* **a**sg (*Sc* **a**sg) and *STE4* haploid specific gene (*Sc* hsg) reporter constructs. (E) *S. cerevisiae* α2 regions 1 and 3 were swapped for the aligning sequence in the *C. albicans* **a**2 protein, genome-integrated in a *MATΔ* background, and assayed for repression of the *Sc* **a**sg reporter construct. (F) An array of **a**sg *cis*-regulatory sequences were selected from the *Kluyveromyces* and *Candida* clades based on their distribution across a range of similarity values to the *S. cerevisiae* **a**sg PSSM (Table S3). Purple shading indicates where α2 binds in *S. cerevisiae* and green shading indicates where Mcm1 binds. Yellow text highlights nucleotides that appear in the consensus binding-sites for *S. cerevisiae* α2. (G) PSSM for α2 alone site, **a**2/α2 site, and **a**2 site alone. (H) The *C. albicans RAM2* was mutated at key residues for α2 binding and tested for their ability to support repression. All values reported in bar graphs are a mean (*n*=3) and standard error of the mean. In each phylogenetic tree, the purple circle marks the gain of α2-mediated repression of **a**sgs and the pink X marks the loss of **a**2.

Figure 4

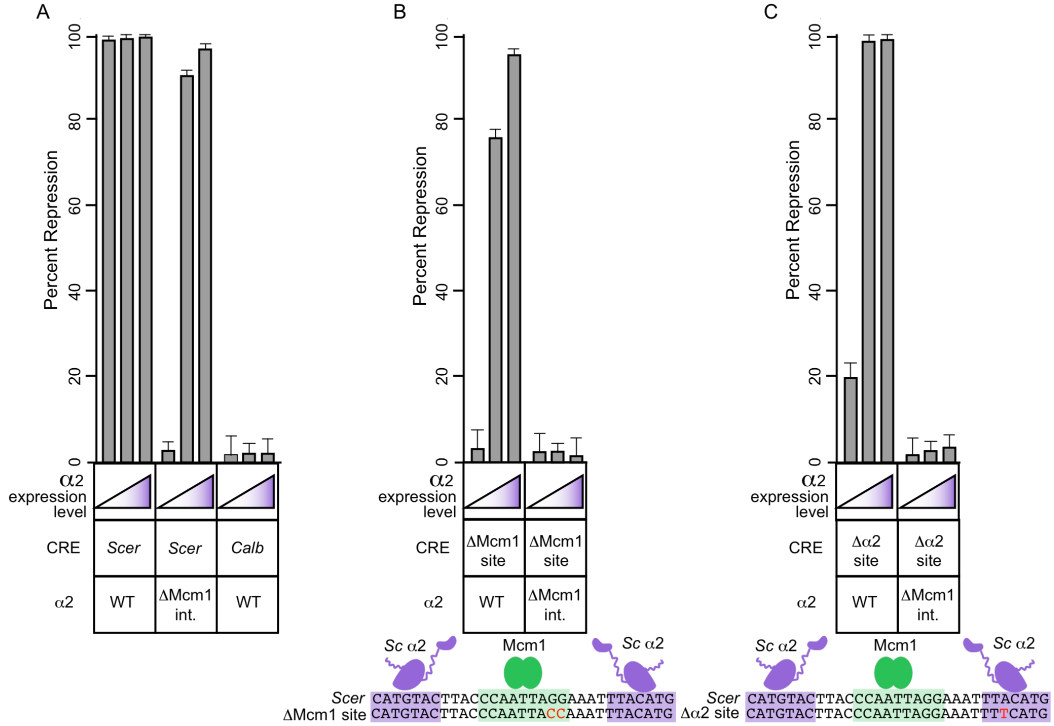


Figure 4: The contribution of non-specific protein interactions to early intermediates

(A) Wild-type *S. cerevisiae* α2 (WT) or mutant *S. cerevisiae* α2 with its Mcm1 interaction region replaced by the aligning sequence from *C. albicans* (ΔMcm1 int.) were tested for the ability to repress the *S. cerevisiae STE2* (*Scer)* or *C. albicans RAM2 (Calb)* **a**sg reporter. The α2 proteins were tested either at the endogenous level, using a strong promoter (*TEF1*), or using a very strong promoter (*TDH3*). (B) Both α2 constructs from (A) were tested for the ability to repress a modified *S. cerevisiae STE2* **a**sg *cis*-regulatory reporter construct where the Mcm1 binding site was compromised (ΔMcm1 site). (C)

Both α2 constructs from (A) were tested for the ability to repress a modified *S. cerevisiae STE2* **a**sg *cis*-regulatory reporter construct where the α2 binding site was compromised (Δα2 site). In all panels, the purple and green shading represents the binding site of α2 and Mcm1, respectively. All values reported in bar graphs are a mean (*n*=3) and standard error of the mean.

Figure 5

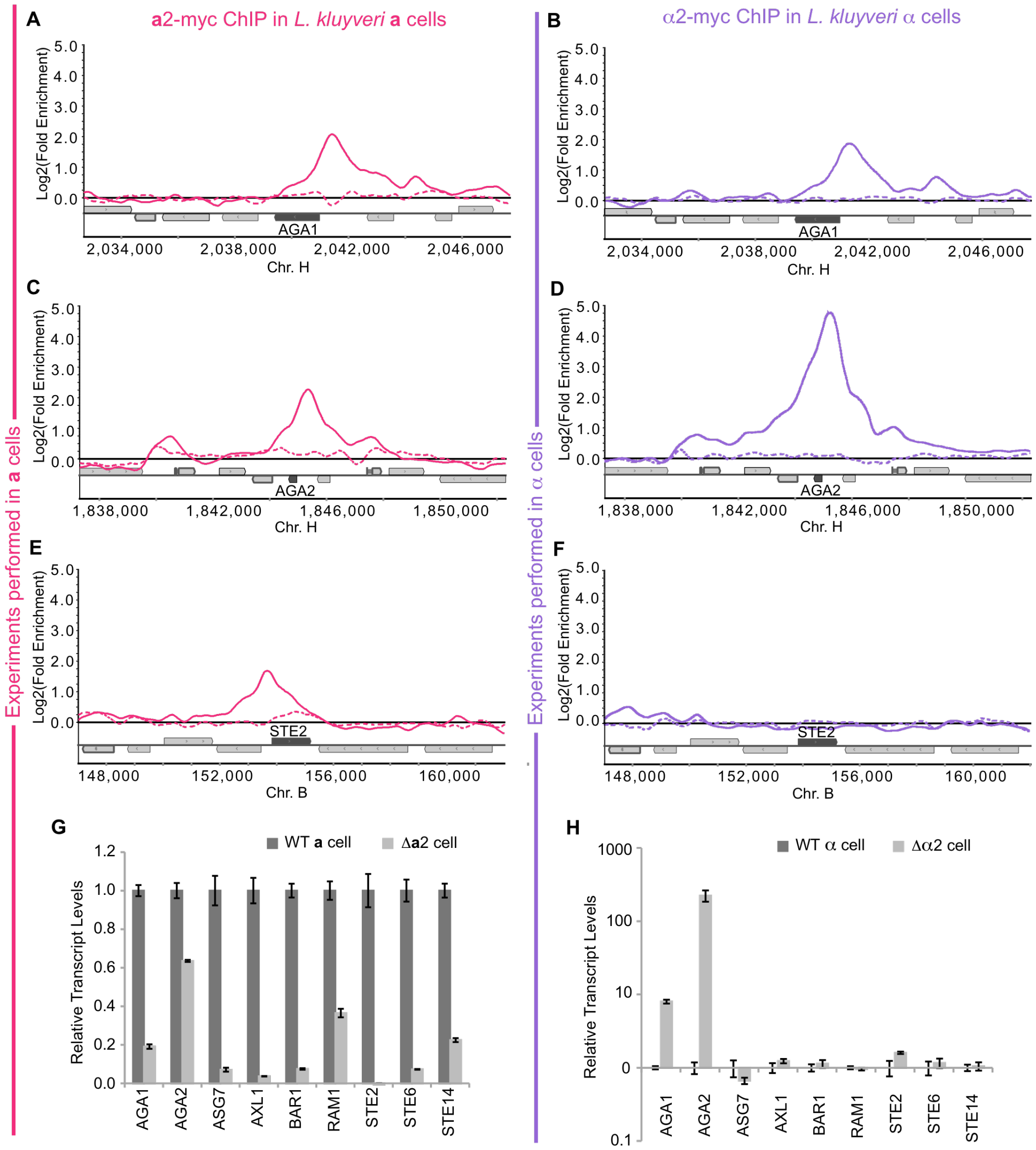


Figure 5: Regulation of the asgs in *Lachancea kluyveri*

(A-F) ChIP-chip was performed using anti-cMyc antibodies in a C-terminal myc-tagged *MATa2* **a** cells (A, C, and E solid, pink lines), wild-type **a** cells (A, C, and E dotted, pink lines), C-terminal myc-tagged MATα2 α cells (B, D, and F solid, purple lines) or wild-type α cells (B, D, and F dotted, purple lines). Wild-type cells serve as untagged controls. ChIP-chip enrichment profiles are shown for *AGA1* (A and B), *AGA2* (C and D) and *STE2* (E and F). Genes (grey rectangles) are displayed below the line if transcribed to the left and above the line if transcribed to the right. (G, H) The transcript levels of the **a**sgs in a wild-type or Δ*MAT**a**2* **a** cell (G) and in a wild-type or Δ*MATα2* α cell (H) were measured relative to *ACT1* by RT-qPCR. The relative transcript abundance for each gene was normalized to the abundance in wild-type **a** cells (G) or in wild-type α cells (H). Displayed is the mean (*n*=3) and standard error of the mean.

Figure 6

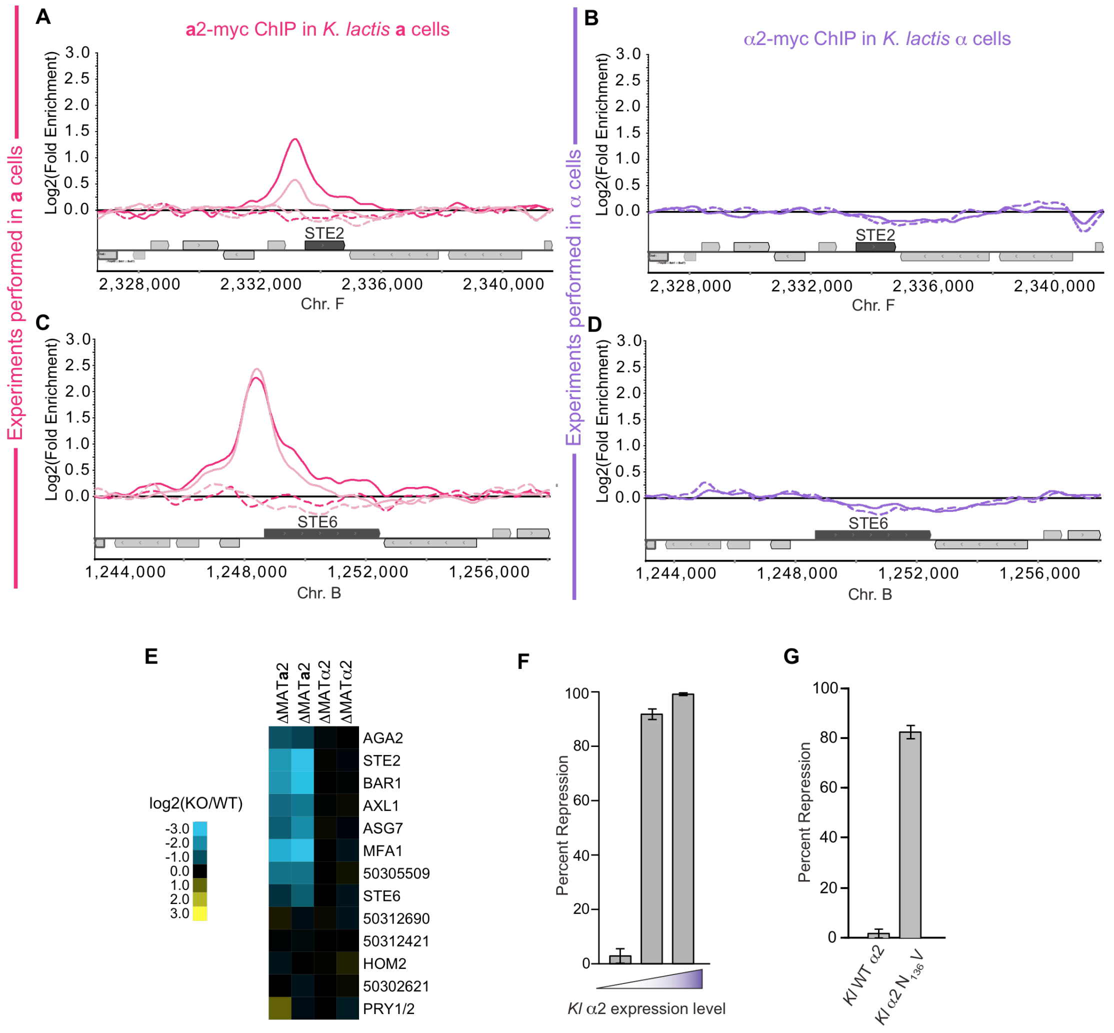
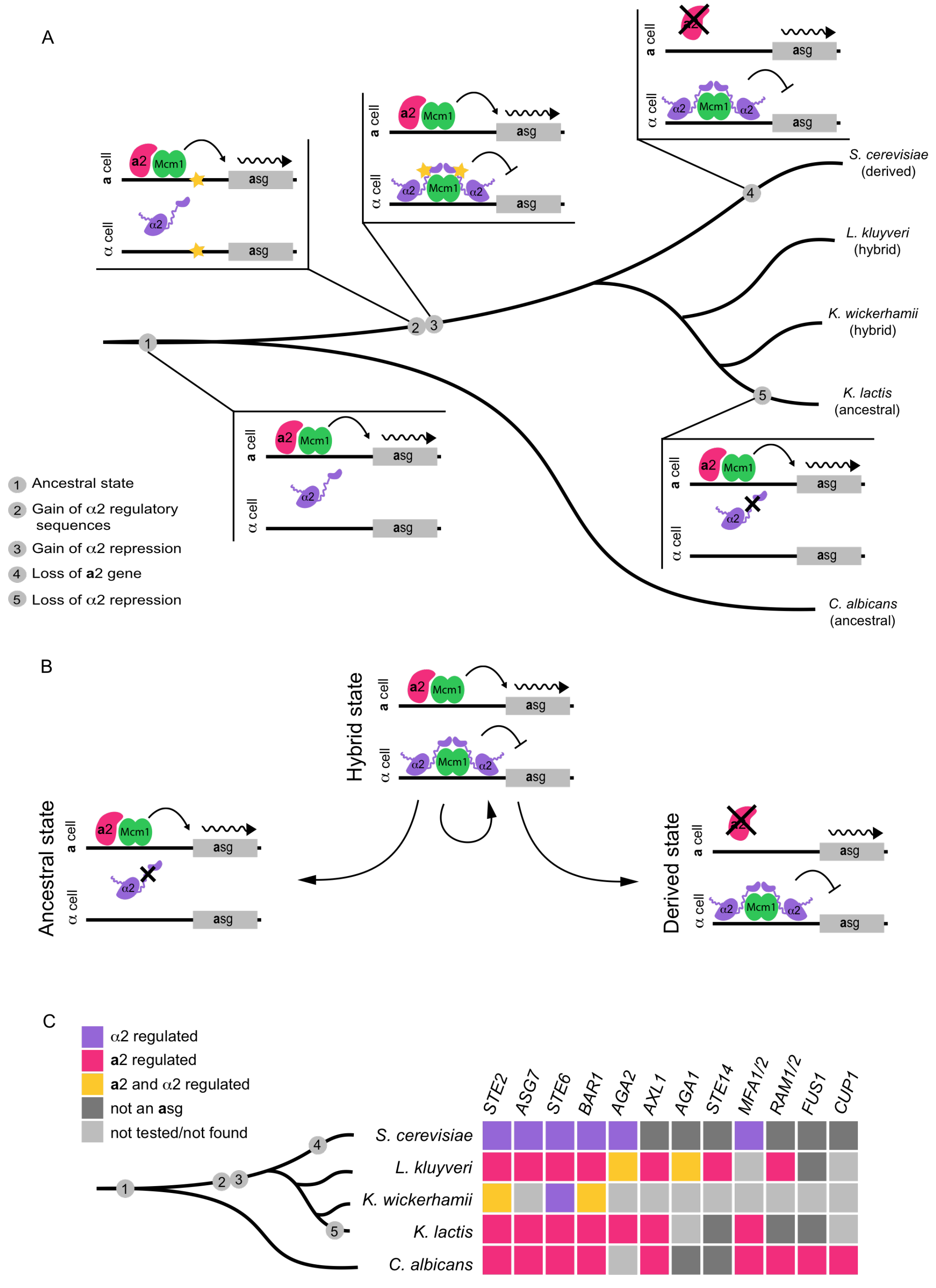


Figure 6: Regulation of the asgs in *Kluyveromyces lactis*

(A-D) ChIP-chip was performed using anti-cMyc antibodies in a C-terminal myc-tagged *MATa2* **a** cells (A and C solid, pink lines), wild-type **a** cells (A and C dotted, pink lines), C-terminal myc-tagged *MATα2* α cells (B and D solid, purple lines) or wild-type α cells (B and D dotted, purple lines). Wild-type cells serve as untagged controls. For ChIP performed in **a** cells (A and C), two conditions were used: one with pheromone induction (dark pink) and one without (light pink). ChIP-chip enrichment profiles are shown for *STE2* (A and B), and *STE6* (C and D). Genes (grey rectangles) are displayed below the line if transcribed to the left and above the line if transcribed to the right. (E) Results for orthologs of the **a**sgs from an expression array comparing mRNA levels from Δ*MAT**a**2* **a** cells to wild-type **a** cells (two left columns) or mRNA levels from Δ*MATα2* α cells to wild-type α cells (two right columns). (F, G) The *K. lactis* α2 protein was assayed for its ability to repress a *S. cerevisiae STE2* operator sequence using a β-gal reporter. (F) Wild-type *K. lactis* α2 was expressed in a *S. cerevisiae MATΔ* cell using promoters of increasing strength. (G) Wild-type *K. lactis* α2 or *K. lactis* α2 with a single point mutation ($N_{136}V$) was expressed in a *S. cerevisiae MATΔ* cell using the endogenous *S. cerevisiae* α2 promoter. Displayed are the mean (*n*=3) and standard error of the mean.

Figure 7



	STE2	ASG7	STE6	BAR1	AGA2	AXL1	AGA1	STE14	MFA1/2	RAM1/2	FUS1	CUP1
S. cerevisiae	🟪	🟪	🟪	🟪	🟪	⬛	⬛	⬛	🟪	⬛	⬛	⬛
L. kluyveri	🟥	🟥	🟥	🟥	🟨	🟥	🟨	🟥	⬜	🟥	⬛	⬜
K. wickerhamii	🟨	⬜	🟪	🟨	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
K. lactis	🟥	🟥	🟥	🟥	🟥	🟥	⬜	⬛	🟥	⬛	⬛	⬜
C. albicans	🟥	🟥	🟥	🟥	⬜	🟥	⬛	⬛	🟥	🟥	🟥	🟥

Figure 7: The gain of the hybrid regulatory state facilitated diversification of asg regulation

(A) The evolutionary trajectory of the gain of repression by α2 is shown for a representative **a**sg. Major evolutionary events are indicated by numbered, grey circles. Gains, either in *cis* or *trans* are indicated by yellow stars and losses by a black "x". The regulatory state of the extant yeast are shown (ancestral indicates **a**2 activation only, derived indicates α2 repression only and hybrid indicates both modes of regulation). (B) The hybrid intermediate can "resolve" in different ways. It can revert to the ancestral mode of regulation through loss of the derived mode (left arrow; *K. lactis*), maintain the hybrid in some fashion (circular, center arrow; *K. wickerhamii* and *L. kluyveri*), or lose the ancestral mode of regulation (right arrow; *S. cerevisiae*). (C) Individual genes are regulated differently between and within species. On the left is a recapitulation of part A of this figure. **a**sgs are listed by the *S. cerevisiae* orthologs on the top of the figure and their mode of regulation (if available) are indicated for each species by a colored square (see key in figure).

References:

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Research *37*, W202-208.

Booth, L.N., Tuch, B.B., and Johnson, A.D. (2010). Intercalation of a new tier of transcription regulation into an ancient circuit. Nature *468*, 959-963.

Butler, G., Kenny, C., Fagan, A., Kurischko, C., Gaillardin, C., and Wolfe, K.H. (2004). Evolution of the MAT locus and its Ho endonuclease in yeast species. Proc Natl Acad Sci USA *101*, 1632-1637.

Carroll, S.B. (2005). Evolution at two levels: on genes and form. PLoS Biol *3*, e245.

Chan, Y.F., Marks, M.E., Jones, F.C., Villarreal, G., Shapiro, M.D., Brady, S.D., Southwick, A.M., Absher, D.M., Grimwood, J., Schmutz, J.*, et al.* (2010). Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. Science *327*, 302-305.

Davidson, E.H., and Erwin, D.H. (2006). Gene regulatory networks and the evolution of animal body plans. Science *311*, 796-800.

Doebley, J., and Lukens, L. (1998). Transcriptional regulators and the evolution of plant form. Plant Cell *10*, 1075-1082.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research *32*, 1792-1797.

Galgoczy, D.J., Cassidy-Stone, A., Llinás, M., O'Rourke, S.M., Herskowitz, I., DeRisi, J.L., and Johnson, A.D. (2004). Genomic dissection of the cell-type-specification circuit in Saccharomyces cerevisiae. Proc Natl Acad Sci USA *101*, 18069-18074.

Gompel, N., Prud'homme, B., Wittkopp, P.J., Kassner, V.A., and Carroll, S.B. (2005). Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. Nature *433*, 481-487.

Guarente, L., and Ptashne, M. (1981). Fusion of Escherichia coli lacZ to the cytochrome c gene of Saccharomyces cerevisiae. Proc Natl Acad Sci USA *78*, 2199-2203.

Henikoff, S., and Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA *89*, 10915-10919.

Herskowitz, I. (1989). A regulatory hierarchy for cell specialization in yeast. Nature *342*, 749-757.

Homann, O.R., and Johnson, A.D. (2010). MochiView: versatile software for genome browsing and DNA motif analysis. BMC biology *8*, 49.

Hull, C.M., and Johnson, A.D. (1999). Identification of a mating type-like locus in the asexual pathogenic yeast Candida albicans. Science *285*, 1271-1275.

Johnson, A.D., and Herskowitz, I. (1985). A repressor (MAT alpha 2 Product) and its operator control expression of a set of cell type specific genes in yeast. Cell *42*, 237-247.

Keleher, C.A., Goutte, C., and Johnson, A.D. (1988). The yeast cell-type-specific repressor alpha 2 acts cooperatively with a non-cell-type-specific protein. Cell *53*, 927-936.

Komachi, K., Redd, M.J., and Johnson, A.D. (1994). The WD repeats of Tup1 interact with the homeo domain protein alpha 2. Genes & Development *8*, 2857-2867.

Lavoie, H., Hogues, H., Mallick, J., Sellam, A., Nantel, A., and Whiteway, M. (2010). Evolutionary tinkering with conserved components of a transcriptional regulatory network. PLoS Biol *8*, e1000329.

Lynch, M. (2007). The evolution of genetic networks by non-adaptive processes. Nat Rev Genet *8*, 803-813.

Lynch, V.J., Leclerc, R.D., May, G., and Wagner, G.P. (2011). Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. Nat Genet *43*, 1154-1159.

Madhani, H. (2007). From a to alpha: Yeast as a Model for Cellular Differentiation, 1 edn (Cold Spring Harbor, NY, Cold Spring Harbor Laboratory Press).

Mead, J., Zhong, H., Acton, T.B., and Vershon, A.K. (1996). The yeast alpha2 and Mcm1 proteins interact through a region similar to a motif found in homeodomain proteins of higher eukaryotes. Mol Cell Biol *16*, 2135-2143.

Smith, D.L., and Johnson, A.D. (1994). Operator-constitutive mutations in a DNA sequence recognized by a yeast homeodomain. EMBO J *13*, 2378-2387.

Strathern, J., Hicks, J., and Herskowitz, I. (1981). Control of cell type in yeast by the mating type locus. The alpha 1-alpha 2 hypothesis. J Mol Biol *147*, 357-372.

Tan, S., and Richmond, T.J. (1998). Crystal structure of the yeast MATalpha2/MCM1/DNA ternary complex. Nature *391*, 660-666.

Tanay, A., Regev, A., and Shamir, R. (2005). Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. Proc Natl Acad Sci USA *102*, 7203-7208.

Taylor, J.W., and Berbee, M.L. (2006). Dating divergences in the Fungal Tree of Life: review and new analyses. Mycologia *98*, 838-849.

Thornton, J.W. (2004). Resurrecting ancient genes: experimental analysis of extinct molecules. Nat Rev Genet *5*, 366-375.

Tishkoff, S.A., Reed, F.A., Ranciaro, A., Voight, B.F., Babbitt, C.C., Silverman, J.S., Powell, K., Mortensen, H.M., Hirbo, J.B., Osman, M.*, et al.* (2007). Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet *39*, 31-40.

Tsong, A.E., Miller, M.G., Raisner, R.M., and Johnson, A.D. (2003). Evolution of a combinatorial transcriptional circuit: a case study in yeasts. Cell *115*, 389-399.

Tsong, A.E., Tuch, B.B., Li, H., and Johnson, A.D. (2006). Evolution of alternative transcriptional circuits with identical logic. Nature *443*, 415-420.

Tuch, B.B., Li, H., and Johnson, A.D. (2008). Evolution of eukaryotic transcription circuits. Science *319*, 1797-1799.

Vershon, A.K., and Johnson, A.D. (1993). A short, disordered protein region mediates interactions between the homeodomain of the yeast alpha 2 protein and the MCM1 protein. Cell *72*, 105-112.

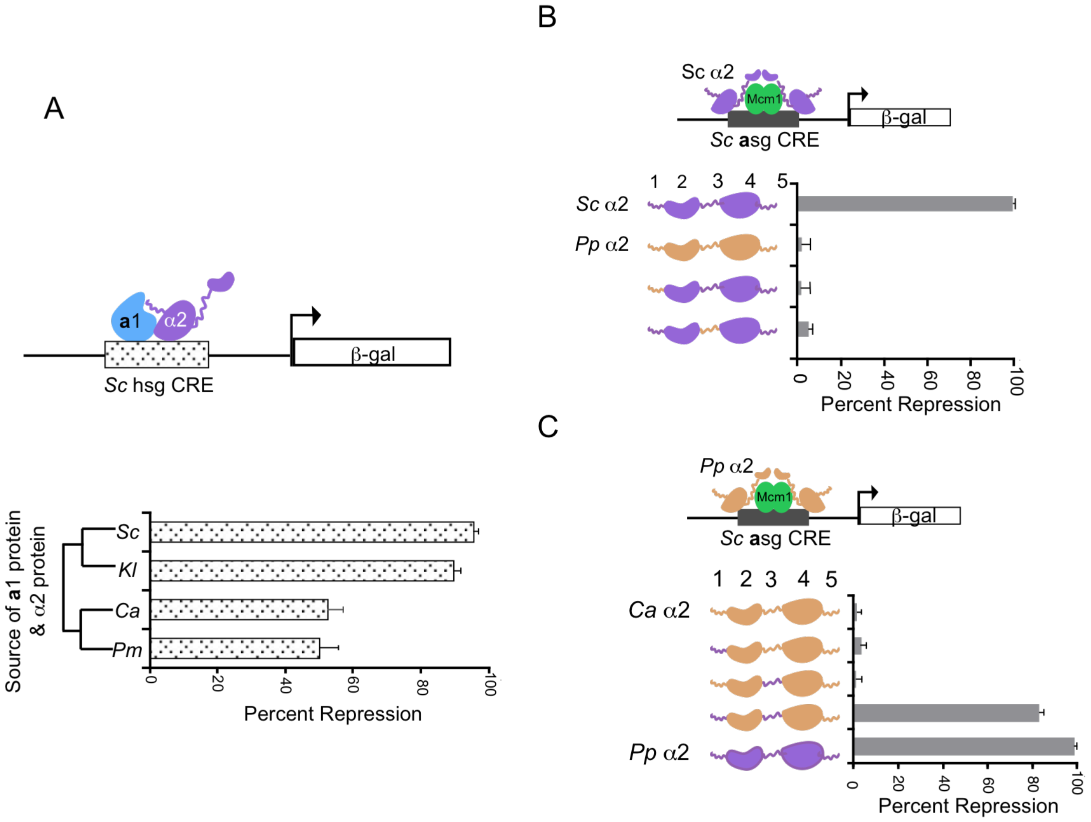
Wagner, G. (2008). Resurrecting the role of transcription factor change in developmental evolution. Evolution.

Wohlbach, D.J., Thompson, D.A., Gasch, A.P., and Regev, A. (2009). From elements to modules: regulatory evolution in Ascomycota fungi. Curr Opin Genet Dev, 1-8.

Wray, G.A. (2007). The evolutionary significance of cis-regulatory mutations. Nat Rev Genet *8*, 206-216.

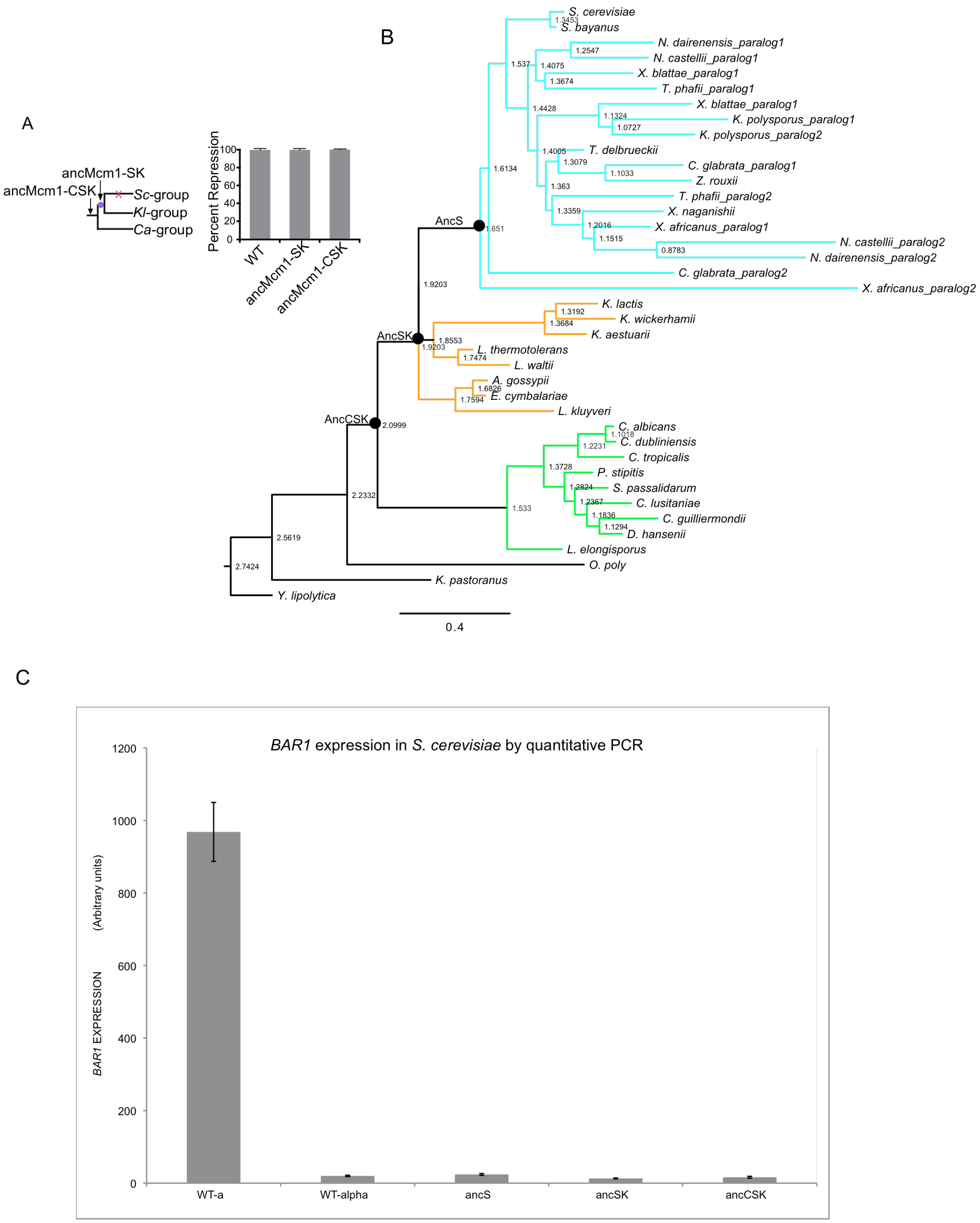
Supplemental Figures

Supplemental Figure 1



Supplemental Figure 1, related to Figures 2 and 3: Further support that the evolution occurred of asg repression occurred in the ancestor of the *Saccharomyces-Kluyveromyces* clade and required the gain of a Tup1 and Mcm1 interaction within the α2 protein

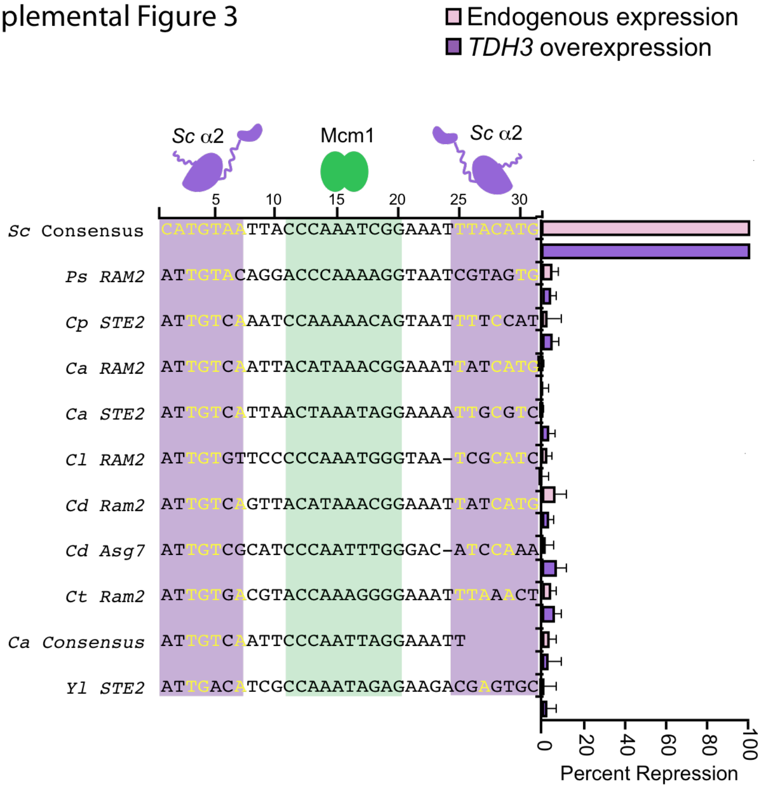
(A) *K. lactis, C. albicans* and *P. membranifacians* failed to repress **a**sg expression (Figure 2B-C), but retains the capacity to repress a haploid-specific gene reporter with a species-matched **a**1 protein in *S. cerevisiae* (the *cis*-regulatory sequence was taken from the *S. cerevisiae STE4* gene). This demonstrated that each α2 protein, although unable to repress the **a**sgs, was functional in *S. cerevisiae*. (B) Region-swapping between *S. cerevisiae* and *Pichia pastoris* α2 protein. Both region 1 (Tup1 interaction) and region 3 (Mcm1 interaction) in *S. cerevisiae* α2 were replaced with the aligning sequence from *P. pastoris*. (**C**), Regions 1 and 3 of the *S. cerevisiae* α2 were swapped into the *P. pastoris* α2 protein. Each construct was genome-integrated using pNH604 in a MATΔ background and assayed for the ability to repress the *S. cerevisiae STE2* **a**-specific gene (*Sc* **a**sg). The *P. pastoris* α2 sequence performed essentially identically to the *C. albicans'* sequence (Figure 3D & E). Values reported in bar graphs are a means (*n*=3) and standard errors of the mean.



Supplemental Figure 2, related to Figure 3: The evolution of α2-repression exploited an ancestral Mcm1 interaction surface

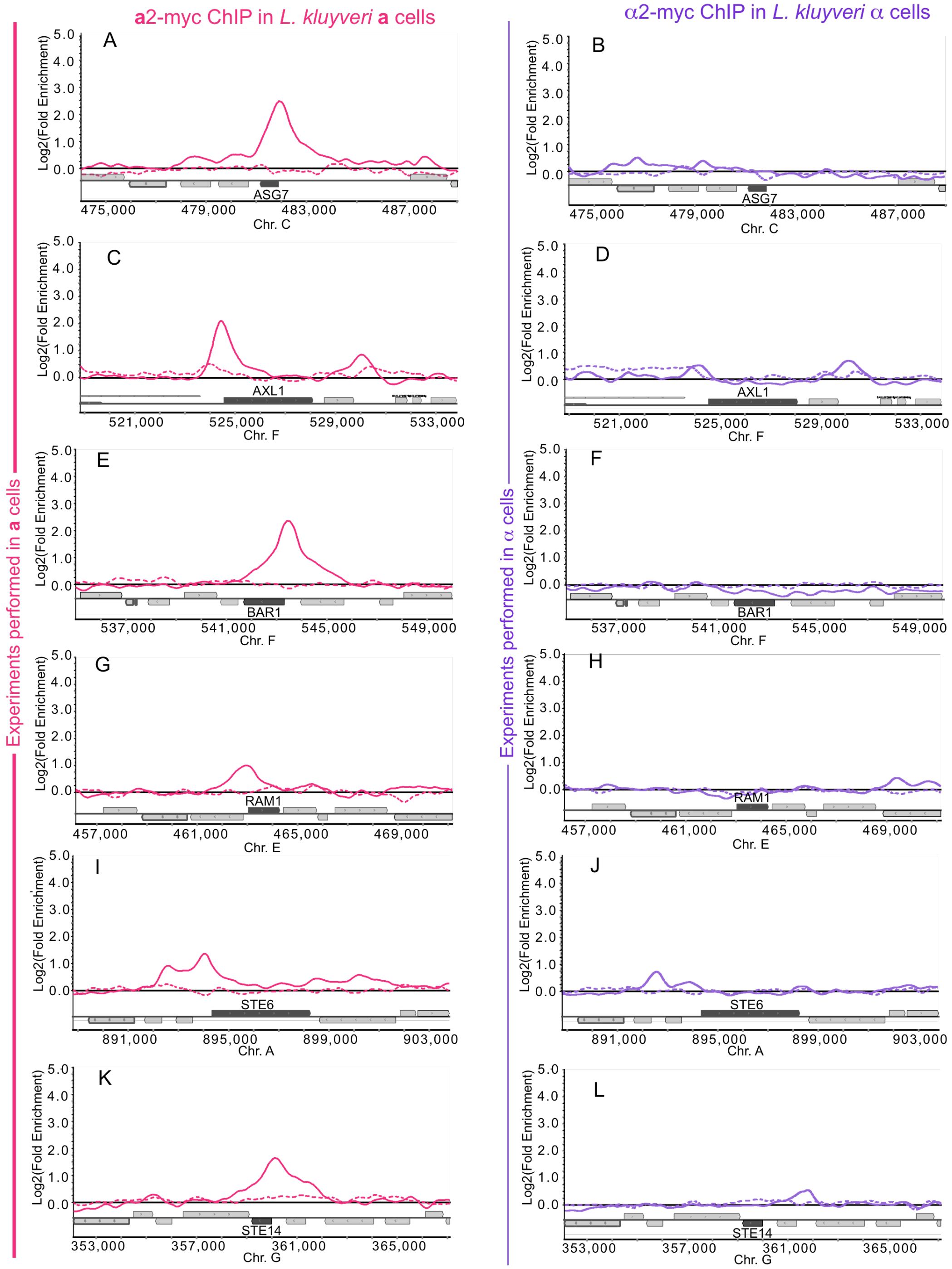
(A) Ancestral Mcm1 proteins complement a deletion of the *S. cerevisiae* Mcm1 protein in repression of the **a**sgs. Two ancestral protein cDNAs were synthesized and used to replace the modern *S. cerevisiae* Mcm1 MADS-box domain. Repression of the *S. cerevisiae* **a**sg reporter was determined by comparing expression levels in **a** cells and α cells. (B) The gene tree used as a guide for the ancestral reconstruction (see also Table S1) was built using PhylML (Guindon and Gascuel 2003). Branch node confidence was calculated with approximate likelihood ratios (shown at branch nodes). AncKS represents the last common ancestor of both the *Saccharomyces* and *Kluyveromyces* groups (blue and pink, respectively). AncCSK also includes the *Candida* clade sequences (green). The AncS (*Saccharomyces* ancestral Mcm1, blue) was not tested. (C) Verificiation of the results from the **a**sg reporter (panel A) by assaying repression of the endogenous *S. cerevisiae* **a**sgs through gene expression profiling by RT-qPCR of the *BAR1* transcript. Expression levels were quantified in *S. cerevisiae* MAT**a** and MATα cells, as well in *S. cerevisiae* α cells where the endogenous Mcm1 MADS-box domain was replaced with the ancestral MADS-box domains. Values reported in bar graphs are means (*n*=3) and standard errors of the mean. These repression results were also confirmed by microarray in the strains that expressed the ancestrally reconstructed Mcm1 proteins (data not shown).

Supplemental Figure 3



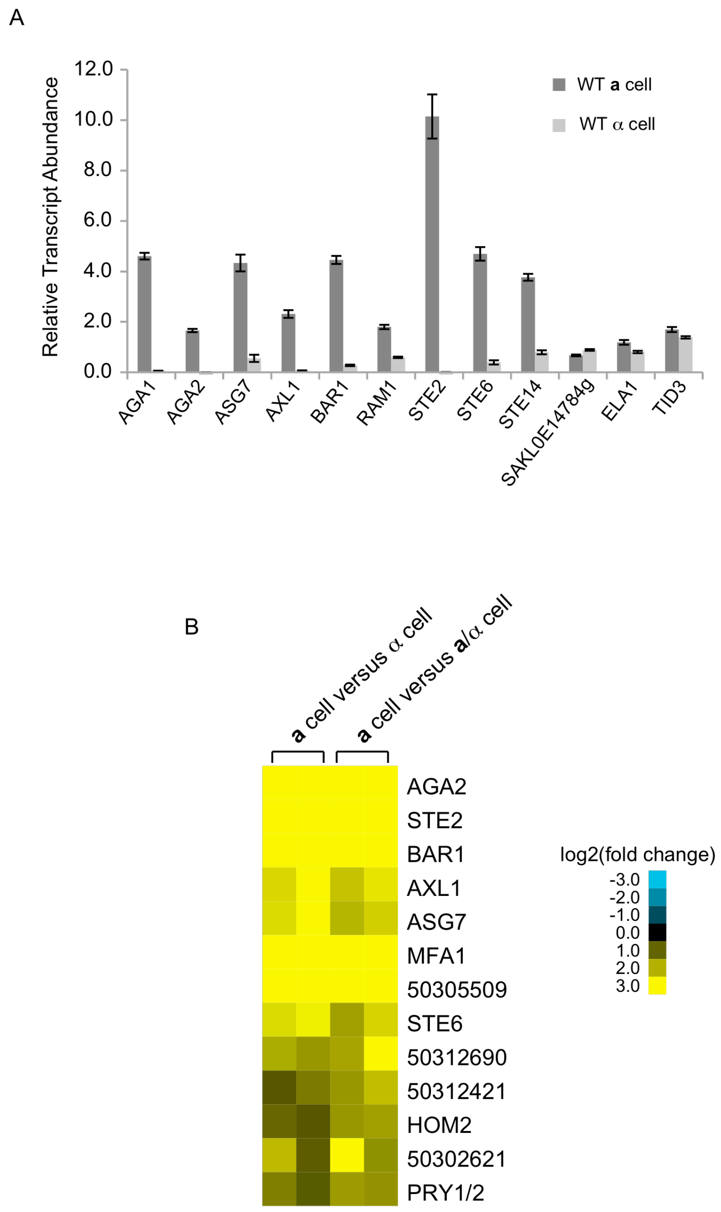
Supplemental Figure 3, related to Figures 1 and 3: Over-expression of *S. cerevisiae* α2 failed to repress out-group a-specific gene *cis*-regulatory sequences

Over-expressing *S. cerevisiae* α2 from the *TDH3* promoter fails to repress the *Candida* group and *Y. lipolytica* **a**-specific gene *cis*-regulatory sequences, indicating that despite the close resemblance between ancestral regulatory sequences and those repressed by α2 in the derived regulatory state, modifications needed to occur to the ancestral **a**-specific gene *cis*-regulatory sequences early in the evolution of this new regulatory mode to support repression in intermediates. Values reported in bar graphs are means (*n*=3) and standard errors of the mean.



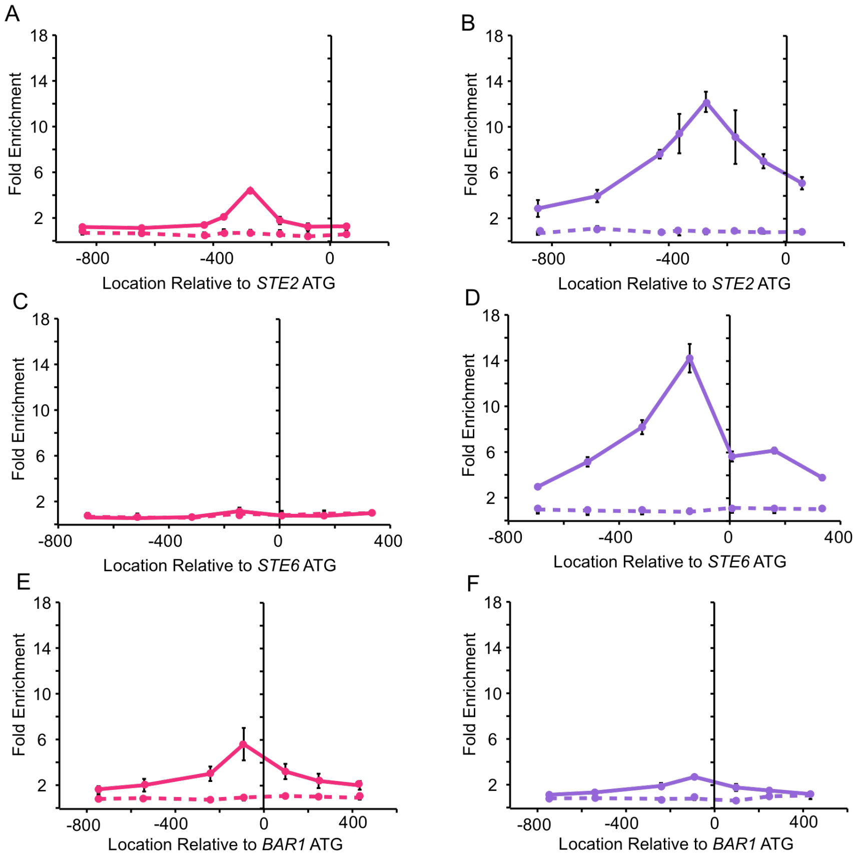
Supplemental Figure 4, related to Figure 5: Regulation of *L. kluyveri* a-specific genes

ChIP-chip was performed using anti-cMyc antibodies in C-terminal myc-tagged MAT**a**2 **a** cells (A, C, E, F, G, I , and K solid, pink lines), wild-type **a** cells (A, C, E, F, G, I , and K dotted, pink lines), C-terminal myc-tagged MATα2 α cells (B, D, F, H, J, and L solid, purple lines) or wild-type α cells (B, D, F, H, J, and L dotted, purple lines). Wild-type cells serve as untagged controls. ChIP-chip enrichment profiles are shown for *ASG7* (A and B), *AAL1* (C and D), *BAR1* (E and F), *RAM1* (G and H), *STE6* (I and J), and *STE14* (K and L). Genes (grey rectangles) are displayed below the line if transcribed to the left and above the line if transcribed to the right. Data is visualized with MochiView (Homann and Johnson 2010).



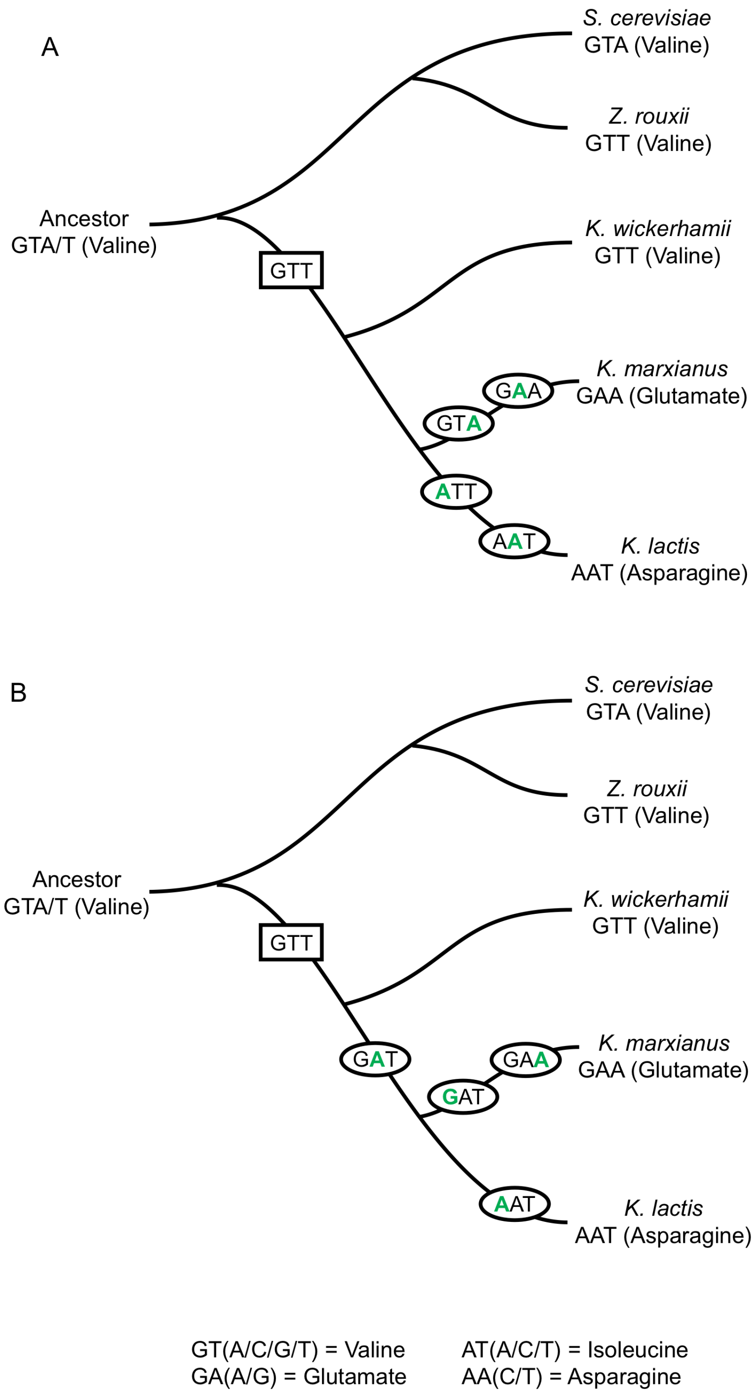
Supplemental Figure 5, related to Figure 5: The *K. lactis* and *L. kluyveri* a-specific genes

RNA was isolated and reverse transcribed from wild-type cells. (A) Transcript levels of **a**-specific gene orthologs in *L. kluyveri* **a** and α cells were quantified and measured relative to *ACT1* by RT-qPCR. The mean (*n*=3) and standard error of the mean are shown. (B) Gene expression arrays were used to measure transcript abundance in wild-type *K. lactis* **a**, α, and **a**/α cells. Three replicates were performed. Shown are the genes which were up at least 2-fold in **a** cells versus α and **a**/α cells.



Supplemental Figure 6, related to Figure 7: Regulation of *K. wickerhamii* a-specific genes

ChIP-quantitative PCR was performed using anti-HA antibodies in a N-terminal HA-tagged MAT**a**2 cells (A, C, and E solid, pink lines), C-terminal myc-tagged MATα2 (B, D, and F solid, purple lines) or wild-type cells (A-F dotted lines). Wild-type cells serve as untagged controls. Since *K. wickerhamii* can undergo mating type switching and isolation of pure populations of a single mating type was not possible, we overexpressed either the **a**2 or α2 protein from constructs integrated in their endogenous loci to increase the likelihood of detecting binding in a mixed cell population. *BAR1*, *STE2*, and *STE6* had clear binding enrichments for at least one of the regulators. Using primer sets that tiled across the upstream regions of each of these three genes, we measured the binding of the two regulators in finer detail. ChIP-qPCR enrichment profiles are shown for *STE2* (A and B), *STE6* (C and D) and *BAR1* (E and F). The start codon of each gene is at the 0-coordinate on the x-axis and transcription/translation proceeds to the right. Each point represents the mean enrichment (relative to a region of the *ACT1* promoter) of 3 replicates and error bars show standard error of the mean. We note that with any ChIP experiment, but in particular with a mixed cell population, the lack of an observed binding event does not necessarily indicate that a gene is not bound *in vivo*.



Supplemental Figure 7, related to Figure 6: The mutational path of the loss of α2 repression in the *Kluyveromyces*

Presented are two possible mutational paths towards the single amino acid mutations in *K. lactis* (asparagine, position 136) and *K. marxianus* (glutamate) α2 proteins. In the ancestor of the *Kluyveromyces* and *Saccharomyces* lineages this residue is a valine (likely encoded by either a GTA codon or a GTT codon). We show only a GTT ancestor but note that these mutational paths are also accessible by an ancestor with a GTA codon if a mutation converts it to a GTT codon first. We considered the only permissible mutational paths to be those that pass through the four amino acids observed at this position in the *Kluyveromyces* and *Saccharomyces* clades (valine, isoleucine, glutamate and asparagine). Although other mutational paths (including others that only pass through these four amino acids) are possible, the two presented here require a minimal number of mutations (4).

Extended Experimental Procedures

Identification and analysis of ***a****-specific gene cis-regulatory sequences*

The **a**-specific genes have been identified in several yeast species by gene expression analysis (*S. cerevisiae* and *C. albicans* (Galgoczy, Cassidy-Stone et al. 2004) (Tsong, Miller et al. 2003) (Tsong, Tuch et al. 2006) and *K. lactis* and *L. kluyveri* (this work)). To identify orthologs of these genes in other genome-sequenced yeasts, we used the experimentally defined **a**-specific gene set from *S. cerevisiae* and *C. albicans* as a seed sequence set and utilized tBLASTN to search 32 additional hemiascomycete genomes available on the National Center for Biotechnology Information (NCBI) website (http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi?organism=fungi) for orthologs to these genes. The best-match hit sequence for each search was then reciprocally BLASTed against the *S. cerevisiae* genome to eliminate false positives. If the hit sequence held greater similarity to a protein sequence in *S. cerevisiae* other than the seed sequence, it was eliminated from the data set as false positive. Additionally, we defined the 600bp upstream of each ortholog as the promoter sequence.

To identify **a**-specific gene *cis*-regulatory sequences across the hemiascomcyetes, our **a**-specific gene promoter lists were submitted to MEME (Bailey, Boden et al. 2009). To define a PSSM for the derived (repression) regulatory mode, we submitted the promoter sequences from *S. cerevisiae* and its close relatives *S. mikatae, S. paradoxus,* and *S. bayanus* to MEME (Table S3). To define a PSSM for the ancestral (activation) regulatory mode, we submitted the promoter sequences from *C. albicans* and its close relatives *C. dubliniensis,* and *C. tropicalis* (Table S3). The derived and ancestral PSSM were then used to scan other hemiascomycete **a**-specific gene promoter sequence sets using the MEME utility MAST (Table S2) (Bailey, Boden et al. 2009). Sequences returned by this analysis with E-values below 1.0 (for either the ancestral or derived PSSM search) were then defined as **a**-specific gene *cis*-regulatory sequences. The E-values output by MAST also guided the selection of a range of statistical matches to the derived PSSM for the study of the *cis*-regulatory evolution necessary to the gain of α2 repression (Figure 3F).

Ancestral Reconstruction of Mcm1 MADS-box domains

Orthologs of S. cerevisiae Mcm1 were defined by a tBLASTN search of the 35 additional hemiascomyete genomes available at the NCBI website, as well as a BLASTp search of the collection of yeast genomes available only on the Yeast Gene Order Browser webpage (Gordon, Armisén et al. 2011) (http://wolfe.gen.tcd.ie/ygob/). To eliminate false positives, hit sequences were reverse BLASTed against the *S. cerevisiae* genome. A protein alignment was built from this collection of sequences through submission to PRANK (Löytynoja and Goldman 2005). A gene tree based the Mcm1 protein sequence was generated by PhyML (Figure S2B) (Guindon and Gascuel 2003) .Specific parameters used within PhyML that differ from default settings included: use of the JTT model for amino-acid substitutions, the calculation of approximate likelihood ratios (yes/ aL), and the tree topology model selection setting was set to choose a best fit between NNI and SPR models.

A second gene tree was generated from the same protein sequence alignment and the same settings, except without approximate likelihood ratios, for submission to Lazarus for the reconstruction of ancestral sequences (Hanson-Smith, Kolaczkowski et al.

2010). Beyond the default parameters, we chose to set the amino-acid substitution model to JTT, ancestral gaps sequence were allowed, and the *Y. lipolytica* Mcm1 sequence was set as the out-group. All amino-acid positions within the MADS-box domain were reconstructed to certainties of greater than 90% at our branchpoints of interest. Thus, based on established protocol within the field, we did not opt to build alternative ancestral sequences (Finnigan, Hanson-Smith et al. 2012).

Quantification of conservation scores within α2

Orthologs to the *S. cerevisiae* α2 protein were identified by tBLASTN using the additional 35 hemiascomyete genomes available on the NCBI server. False positives were eliminated by reciprocal-BLAST against the *S. cerevisiae* genome. We added the *K. marxiansus* (GenBank AJ617308.1) and *A. gossypii* α2 (Personal communication from Peter Philippsen) to this sequence set independently. This protein sequence set was then aligned using MUSCLE (Edgar 2004).

To quantify conservation scores for the α2 proteins before after the gain of **a**-specific gene repression, the α2 orthologs were divided into two groups depending on whether the parent species branches within the *Saccharomyces-Kluyveromyces* clade (*K. marxianus*, *K. lactis*, *K. aestuarii*, *L. thermotolerans*, *L. waltii*, *K. wickerhamii*, *A. gossypii*, *Z. rouxii*, *S. cerevisiae*, *S. bayanus*, *K. polysporus*, and *K. delphensis*) or the *Candida* clade (*P. membranifaciens*, *P. angusta*, *S. passalidarum*, *C. tropicalis*, *C. dubliniensis*, *C. albicans*, and *P. pastoris*). The evolutionary space spanned by the *Saccharomyces-Kluyveromcyes* clades and the *Candida* clade are estimated to be similar (Taylor and Berbee 2006). Species closely related to *S. cerevisiae* (such as *S. paradoxus,* *S. mikatae,* and *S. bayanus*) were removed from this analysis to normalize the levels of conservation between *S. cerevisiae* α2 scores and *C. albicans* scores. Average conservation scores for each position in the alignment were quantified between the *S. cerevisiae* or the *C. albicans* sequences and the other species in their clades using identity or an amino acid transition matrix. We found the results to be identical whether we used the BLOSUM62 matrix, PAM250 matrix (Figure 3B), or percent identity (Henikoff and Henikoff 1992). The displayed curve in Figure 3B is smoothed by averaging the conservation score at each position with the adjacent 2 amino-residues on other side.

Strain Construction

The α2 orthologs were PCR amplified from the genomic DNA of the species of interest in 100 μL reactions using the New England Biolabs (NEB) Phusion PCR polymerase and accompanying High-Fidelty PCR buffer (PCR reaction conditions followed the NEB Phusion PCR recommendations). These sequences were either fused to the *S. cerevisiae* α2 promoter by fusion PCR (again using the Phusion polymerase and High-Fidelity buffer) and cloned into the pNH605 plasmid (using PspOMI/BamHI) or cloned directly into pNH605 vectors modified to include the TEF (pNH605-TEF) or TDH3 (pNH605- TDH3) promoters (using BamHI/SacI). For species, such as *L. kluyveri,* that include an intron within the α2 gene, we prepared cDNA from α cells in order to PCR amplify the α2 coding sequence without an intron (protocol for prepping cDNA described in RT-qPCR section of Methods). The CUG-codon within *C. albicans* species codes for a serine, instead of leucine as it does among all other fungi (Santos and Tuite 1995). Due to the presence of these CUG codons in *C. albicans* α2, we had the gene codon-optimized by DNA 2.0 for expression in *S. cerevisiae*. All constructs were sequenced to check for mutations within the α2 coding sequence and promoter. Once linearized (by cutting with PmeI), the pNH605 vector integrated into the *S. cerevisiae* genome at the *LUE2* locus in single copy.

The α2 MUSCLE protein sequence alignment was used to guide the design of the genetic swap experiments along with a combination of genetic and structural information (Komachi, Redd et al. 1994) (Vershon and Johnson 1993) (Johnson and Herskowitz 1985; Johnson, Swanson et al. 1998) (Tan and Richmond 1998) (Li, Jin et al. 1998). Regions 1-5 were defined based on the S. cerevisiae α2 protein (Region 1: amino acid 1-21, Region 2: amino acid 22-108, Region 3: amino acid 109-127, Region 4: amino acid 128-188, Region 5: amino acid 189-210). To minimize the risk of truncating a region of the S. cerevisiae α2 sequence within a secondary structural feature, we predicted the positions of α-helices within the un-crystallized portion of the protein using PHYRE (Kelley and Sternberg 2009)). The swap constructs were built through fusion PCR in the manner described earlier and cloned into the pNH605 vectors. For the reverse swap experiments into *C. albicans* α2, the regions aligning to S. cerevisiae α2 region 1 (1-21) and region 3 (109-127) were replaced within the codon-optimized C. albicans protein coding sequence using fusion PCR. All constructs were sequenced to check for mutations.

Site-directed mutagenesis of the α2 coding sequence was performed using the Agilent QuikChange reaction kit following the manual with no modifications.

Ancestral Mcm1 MADS-box domains were codon-optimized for expression in *S. cerevisiae* by DNA 2.0. Ancestral sequences were then fused to the *S. cerevisiae* Mcm1 sequence flanking the endogenous MADS-box domain (fusion PCRs performed as described above). This flanking sequence extended beyond the coding sequence to include 100bp upstream and downstream of the *S. cerevisiae* coding sequence. The 100bp upstream was included to aid in targeting this construct to the Mcm1 locus and the downstream 100bp were included to preserve the endogenous 3' UTR. This fusion construct was cloned into the pFA6a-KanMx6 (Lorenz, Muir et al. 1995) vector using BamHI/SalI. Within this construct, downstream of the kanamycin marker, we also integrated 100bp of homology to the 3' non-coding sequence of the Mcm1 locus beyond Mcm1's 3'-UTR using EcoRI/SacI. This construct was sequenced to check for mutation. By dropping this construct out of the pFA6a-KanMX6 vector using SalI/EcoRI, we were able to replace the endogenous Mcm1 sequence with the ancestral Mcm1 MADS-box constructs by selection on 100 μg/ml kanamycin YPD plates.

To study **a**-specific gene *cis*-regulatory sequences, 25 bp to 40 bp regions centered around the putative regulatory sequence were synthesized as oligonucleotide primers and annealed on a thermocyler using a gradual decline in temperature from 98 °C at -.1 °C/s. Mutated **a**-specific gene *cis*-regulatory were also ordered as oligonucleotide primers, as opposed to using site-directed mutagenesis. NEB Polynucleotide Kinase (PNK) was used to adhere 5' phosphates to these oligonucleotide dimers (reaction run for 1 hour, using NEB recommended amounts of DNA and enzyme). The kinase was killed at 65 °C for 20 minutes and then heated to 98 °C and slow cooled at -.1 °C/s to re-anneal any oligonucleotides that dissociated from a complement during the enzyme treatment. These **a**-specific gene *cis*-regulatory sequences were cloned into a modified version of the Cyc1 reporter construct pLG699z (Guarente and Ptashne 1981) using XhoI.

To monitor repression, we removed one of the XhoI sites (upstream of the endogenous Cyc1 *cis*-regulatory activation sequence) from pLG699z using a digestion with SmaI that cleaved at two sites flanking this upstream XhoI site. The remaining XhoI site (used to clone the **a**-specific gene *cis*-regulatory sequences) was downstream of the endogenous Cyc1 promoter activation sequence and therefore, the inserted repressor sequence will reside in between the transcription start site and the upstream activation sequence.

For all non-directional cloning reactions, the vector was treated with calf intestinal alkaline phosphatase (CIP) from NEB to prevent self-annealing. For all ligation reactions (both with coding sequence constructs and these *cis*-regulatory sequence constructs), we utilized the DNA ligase Fast-Link from Epicentre Biotechnologies. Reactions were performed at recommended DNA and enzyme concentrations. Transformations were performed with chemically competent *E. coli* DH5α cells. The successful insertion of **a-**specific gene *cis*-regulatory sequences within our pLG669z-derivatives was confirmed by PCR and sequencing.

S. cerevisiae strains were generated using a standard lithium acetate transformation in the W303 background. Experiments in which only *cis*-regulatory sequences were tested were performed in either MAT**a** or MATα *S. cerevisiae* strains. Experiments in which the α2 coding sequences were tested were performed in an MATΔ *S. cerevisiae* strain (Galgoczy, Cassidy-Stone et al. 2004).

Gene disruption cassettes for knock-outs and tagging in *K. lactis, L. kluyveri* and *K. wickerhamii* were generated by fusion PCR (Wach 1996). Fusion PCRs were performed in a 50 μL reaction containing 0.5 μL ExTaq (Takara Bio Inc.), 0.25 mM dNTPs, 0.2 μM each primer and approximately 25 ng template. The reactions were incubated as follows: 94° 3:00, [94° 0:30, 50-55° (depending on primer) 0:30, 72° 1:00/kb] x 35, 72° 5:00. The first round of PCR reactions consisted of 3 reactions that amplified the flanking homologous sequence from genomic DNA using primer 1 and 3 or 4 and 6 and amplified the markers from the appropriate plasmids using primers 2 and 5. The KAN marker and the C-terminal myc-tagging marker were amplified from pFA6a-13Myc-kanMX6 (Longtine, McKenzie et al. 1998) and the 3xHA tagging cassette and pTEF promoter from pYMN-20 (Janke, Magiera et al. 2004). The products were purified with the QIAquick PCR Purification Kit (QIAGEN). The second round of amplification (the fusion round) used 1μL of each purified flank PCR product and 2μL of the purified marker PCR product. This product was purified with the QIAquick PCR Purification Kit (QIAGEN).

The purified fusion PCR products were transformed into *K. lactis, K. wickerhammi* and *L. kluyveri* by electroporation (Gojkovic, Jahnke et al. 2000). Transformants were confirmed to be correct by colony PCR using the check primers listed in Table S6. Tagged genes were also verified by sequencing.

Percent Repression

Percent repression was calculated by dividing a matched strain lacking the α2 construct by the strain with the α2 construct. The quotient of this division was then subtracted from one and multiplied by 100 to transform into percentage. For example, the reporter construct containing the *L. kluyveri AGA2 cis*-regulatory sequence had an average strength of ~400 βgal units in a MATΔ background. When *L. kluyveri* α2 is introduced into this strain, the βgal units drop consistently to less than one. One divided by 400 gives a very small number, which then results in a percent repression close to 100%.

RT-quantitative PCR

Yeast were grown in YEPD to OD_{600} = 0.8 (*K. wickerhamii* and *L. kluyveri*) or were phosphate starved (*K. lactis*) as described previously (Tuch, Galgoczy et al. 2008) and then centrifuged at 4000 rpm for 5 minutes. The supernatant removed and pellets frozen in liquid nitrogen. RNA was isolated and reverse transcribed (using SuperScript II) as previously described (Mitrovich, Tuch et al. 2007) with all volumes scaled appropriately. cDNAs were quantified with a Bio-Rad CFX96 Real Time machine in a standard 25 μL reaction using Sybr green and primer sequences are listed in Table S6.

Gene expression arrays

Using a previously designed probe set (Booth, Tuch et al. 2010), *K. lactis* arrays were printed by Agilent using the 8 x 15K format.

K. lactis strains were grown in phosphate starvation media as described previously (Tuch, Galgoczy et al. 2008). The 50 mL cultures were centrifuged for 5 minutes at 4,000 rpm, the pellet resuspended in 10 mL of 1x TE, and centrifuged again. The supernatant was removed and pellets frozen in liquid nitrogen and stored at -80°. RNA was isolated and reverse transcribed as previously described (Mitrovich, Tuch et al. 2007) with the exception that the RNA isolation protocol was scaled to 50 mL cultures and that SuperScript II (Invitrogen) was used.

2 μg of each mutant strain's cDNA or 4μg of each WT strain's cDNA was dried and resuspended in 5 μL 0.1 M sodium bicarbonate. An equivalent volume of Cy3 or Cy5 dye (Amersham) was added (dyes were resuspended in 60 μL of DMSO) and the reaction incubated at 60° for 45 minutes in the dark. Labeled cDNAs were purified using a Clean and Concentrator -5 kit (Zymo Research). For the first biological replicate, WT cDNA was labeled with Cy3 and knock-out cDNA with Cy5. A dye flip was used for the second biological replicate.

0.5μg of the Cy3 and Cy5 labeled cDNA pairs (WT and knock-out) were hybridized to the array overnight, as described in the Agilent protocol. Following hybridization, the arrays were washed as specified by Agilent with the omission of the final wash (acetonitrile with cynide). Arrays were scanned at 5 μm, averaging 2 lines, with an Axon GenePix 4000A scanner. Arrays were gridded using GenePix Pro version 5.1. Global Lowess normalization analysis was performed for each array using a Goulphar script (Dufour, Wesenberg et al. 2010) (The R Foundation for Statistical Computing). Normalized data was collapsed first by averaging the result for all duplicate probes and finally by taking the median of the probes for each ORF. Data was transformed as described for each experiment. Normalized, transformed data can be found in Table S7. Microarray data were clustered using Cluster version 3.0 (de Hoon, Imoto et al. 2004) and visualized using Java TreeView Version 1.1.3 (Saldanha 2004).

Chromatin immunoprecipitation

Tagged strains were utilized for ChIP (with untagged protein strains used as controls). When possible, the appropriate activity of the tagged regulators was confirmed by monitoring **a**-specific gene expression by RT-qPCR. For *K. wickerhamii*, the mixed cell type population made this step impossible.

K. lactis was grown in phosphate starvation media with or without α-pheromone as described previously (Tuch, Galgoczy et al. 2008). *K. wickerhamii* and *L. kluyveri* were grown in YEPD to OD_{600} = 0.4. The ChIP, DNA amplification, labeling and hybridization were carried out as described previously (Nobile, Nett et al. 2009). For the *K. wickerhamii* **a**2 ChIP, 2 μL of 5 mg/mL mouse anti-HA antibody clone 12CA5 (Roche) was used.

The *K. lactis* tiling arrays used are described (Tuch, Galgoczy et al. 2008). The *L. kluyveri* (previously named *S. kluyveri*) genome was downloaded from Genolevures on 10/07/2010. The *L. kluyveri* tiling array probe set was created using chipD (Dufour, Wesenberg et al. 2010) using Tm model 3 and the default settings with the exception of probe length (minimum probe length = 45, ideal probe length = 60, maximum probe length = 60). Custom *L. kluyveri* tiling arrays were printed by Agilent using the 2 x 105K design. Probe enrichment values for individual or replicate experiments were merged and smoothed using the "Create Smoothed Tiled Set from Data Set(s)" utility in MochiView using the default settings (Homann and Johnson 2010). The identification of statistically significant binding events was performed with MochiView using the "Extract Peaks from Data Set(s)" utility (Homann and Johnson 2010). Peaks were identified from the smoothed and merged tagged strain ChIP using a $\log_2$ cut-off post-smoothing of 0.58 and a p-value less than or equal to 0.001 and the appropriate, untagged control was used as the control data set with $\log_2$ cut-off of 0.27.

For *K. wickerhamii,* ChIP-quantitative PCR was performed following immunoprecipitation. 5μL of a 1/33 dilution of immunoprecipitated DNA (tagged or untagged) or dilution series of whole-cell extracted DNA (as a standard curve) was added to a 20μL, standard, SYBR green qPCR master mix. DNA was quantified on a BioRad CFX96 real-time PCR machine under standard conditions. Primers were designed using the Integrated DNA Technology PrimerQuest utility using the default qPCR settings and are listed in Table S6.

References

Alon, U. (2007). An introduction to systems biology : design principles of biological circuits. Boca Raton, FL, Chapman & Hall/CRC.

Bailey, T. L., M. Boden, et al. (2009). "MEME SUITE: tools for motif discovery and searching." Nucleic Acids Research **37**(Web Server issue): W202-208.

Booth, L. N., B. B. Tuch, et al. (2010). "Intercalation of a new tier of transcription regulation into an ancient circuit." Nature **468**(7326): 959-963.

Butler, G., C. Kenny, et al. (2004). "Evolution of the MAT locus and its Ho endonuclease in yeast species." Proc Natl Acad Sci USA **101**(6): 1632-1637.

Carroll, S. B. (2005). "Evolution at two levels: on genes and form." PLoS Biol **3**(7): e245.

Chan, Y. F., M. E. Marks, et al. (2010). "Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer." Science **327**(5963): 302-305.

Davidson, E. H. and D. H. Erwin (2006). "Gene regulatory networks and the evolution of animal body plans." Science **311**(5762): 796-800.

de Hoon, M. J. L., S. Imoto, et al. (2004). "Open source clustering software." Bioinformatics **20**(9): 1453-1454.

Doebley, J. and L. Lukens (1998). "Transcriptional regulators and the evolution of plant form." Plant Cell **10**(7): 1075-1082.

Dufour, Y. S., G. E. Wesenberg, et al. (2010). "chipD: a web tool to design oligonucleotide probes for high-density tiling arrays." Nucleic Acids Research **38**(Web Server): W321-W325.

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Research **32**(5): 1792-1797.

Finnigan, G. C., V. Hanson-Smith, et al. (2012). "Evolution of increased complexity in a molecular machine." Nature **481**(7381): 360-364.

Galgoczy, D. J., A. Cassidy-Stone, et al. (2004). "Genomic dissection of the cell-type-specification circuit in Saccharomyces cerevisiae." Proc Natl Acad Sci USA **101**(52): 18069-18074.

Gojkovic, Z., K. Jahnke, et al. (2000). "PYD2 encodes 5,6-dihydropyrimidine amidohydrolase, which participates in a novel fungal catabolic pathway." J Mol Biol **295**(4): 1073-1087.

Gompel, N., B. Prud'homme, et al. (2005). "Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila." Nature **433**(7025): 481-487.

Gordon, J. L., D. Armisén, et al. (2011). "Evolutionary erosion of yeast sex chromosomes by mating-type switching accidents." Proc Natl Acad Sci USA **108**(50): 20024-20029.

Guarente, L. and M. Ptashne (1981). "Fusion of Escherichia coli lacZ to the cytochrome c gene of Saccharomyces cerevisiae." Proc Natl Acad Sci USA **78**(4): 2199-2203.

Guindon, S. and O. Gascuel (2003). "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood." Syst Biol **52**(5): 696-704.

Hanson-Smith, V., B. Kolaczkowski, et al. (2010). "Robustness of ancestral sequence reconstruction to phylogenetic uncertainty." Mol Biol Evol **27**(9): 1988-1999.

Henikoff, S. and J. G. Henikoff (1992). "Amino acid substitution matrices from protein blocks." Proc Natl Acad Sci USA **89**(22): 10915-10919.

Herskowitz, I. (1989). "A regulatory hierarchy for cell specialization in yeast." Nature **342**(6251): 749-757.

Homann, O. R. and A. D. Johnson (2010). "MochiView: versatile software for genome browsing and DNA motif analysis." BMC biology **8**(1): 49.

Hull, C. M. and A. D. Johnson (1999). "Identification of a mating type-like locus in the asexual pathogenic yeast Candida albicans." Science **285**(5431): 1271-1275.

Janke, C., M. M. Magiera, et al. (2004). "A versatile toolbox for PCR-based tagging of yeast genes: new fluorescent proteins, more markers and promoter substitution cassettes." Yeast **21**(11): 947-962.

Johnson, A. D. and I. Herskowitz (1985). "A repressor (MAT alpha 2 Product) and its operator control expression of a set of cell type specific genes in yeast." Cell **42**(1): 237-247.

Johnson, P. R., R. Swanson, et al. (1998). "Degradation signal masking by heterodimerization of MATalpha2 and MATa1 blocks their mutual destruction by the ubiquitin-proteasome pathway." Cell **94**(2): 217-227.

Keleher, C. A., C. Goutte, et al. (1988). "The yeast cell-type-specific repressor alpha 2 acts cooperatively with a non-cell-type-specific protein." Cell **53**(6): 927-936.

Kelley, L. A. and M. J. E. Sternberg (2009). "Protein structure prediction on the Web: a case study using the Phyre server." Nat Protoc **4**(3): 363-371.

Komachi, K., M. J. Redd, et al. (1994). "The WD repeats of Tup1 interact with the homeo domain protein alpha 2." Genes & Development **8**(23): 2857-2867.

Lavoie, H., H. Hogues, et al. (2010). "Evolutionary tinkering with conserved components of a transcriptional regulatory network." PLoS Biol **8**(3): e1000329.

Li, T., Y. Jin, et al. (1998). "Crystal structure of the MATa1/MATalpha2 homeodomain heterodimer in complex with DNA containing an A-tract." Nucleic Acids Res **26**(24): 5707-5718.

Longtine, M. S., A. McKenzie, et al. (1998). "Additional modules for versatile and economical PCR-based gene deletion and modification in Saccharomyces cerevisiae." Yeast **14**(10): 953-961.

Lorenz, M. C., R. S. Muir, et al. (1995). "Gene disruption with PCR products in Saccharomyces cerevisiae." Gene **158**(1): 113-117.

Löytynoja, A. and N. Goldman (2005). "An algorithm for progressive multiple alignment of sequences with insertions." Proc Natl Acad Sci USA **102**(30): 10557-10562.

Lynch, M. (2007). "The evolution of genetic networks by non-adaptive processes." Nat Rev Genet **8**(10): 803-813.

Lynch, V. J., R. D. Leclerc, et al. (2011). "Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals." Nat Genet **43**(11): 1154-1159.

Madhani, H. (2007). From a to alpha: Yeast as a Model for Cellular Differentiation. Cold Spring Harbor, NY, Cold Spring Harbor Laboratory Press.

Mead, J., H. Zhong, et al. (1996). "The yeast alpha2 and Mcm1 proteins interact through a region similar to a motif found in homeodomain proteins of higher eukaryotes." Mol Cell Biol **16**(5): 2135-2143.

Mitrovich, Q. M., B. B. Tuch, et al. (2007). "Computational and experimental approaches double the number of known introns in the pathogenic yeast Candida albicans." Genome Res **17**(4): 492-502.

Nobile, C. J., J. E. Nett, et al. (2009). "Biofilm matrix regulation by Candida albicans Zap1." PLoS Biol **7**(6): e1000133.

Rupp, S. (2002). "*LacZ* assays in yeast." Methods in Enzymology **350**: 112-131.

Saldanha, A. J. (2004). "Java Treeview--extensible visualization of microarray data." Bioinformatics **20**(17): 3246-3248.

Santos, M. A. and M. F. Tuite (1995). "The CUG codon is decoded in vivo as serine and not leucine in Candida albicans." Nucleic Acids Research **23**(9): 1481-1486.

Smith, D. L. and A. D. Johnson (1994). "Operator-constitutive mutations in a DNA sequence recognized by a yeast homeodomain." EMBO J **13**(10): 2378-2387.

Strathern, J., J. Hicks, et al. (1981). "Control of cell type in yeast by the mating type locus. The alpha 1-alpha 2 hypothesis." J Mol Biol **147**(3): 357-372.

Tan, S. and T. J. Richmond (1998). "Crystal structure of the yeast MATalpha2/MCM1/DNA ternary complex." Nature **391**(6668): 660-666.

Tanay, A., A. Regev, et al. (2005). "Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast." Proc Natl Acad Sci USA **102**(20): 7203-7208.

Taylor, J. W. and M. L. Berbee (2006). "Dating divergences in the Fungal Tree of Life: review and new analyses." Mycologia **98**(6): 838-849.

Thornton, J. W. (2004). "Resurrecting ancient genes: experimental analysis of extinct molecules." Nat Rev Genet **5**(5): 366-375.

Tishkoff, S. A., F. A. Reed, et al. (2007). "Convergent adaptation of human lactase persistence in Africa and Europe." Nat Genet **39**(1): 31-40.

Tsong, A. E., M. G. Miller, et al. (2003). "Evolution of a combinatorial transcriptional circuit: a case study in yeasts." Cell **115**(4): 389-399.

Tsong, A. E., B. B. Tuch, et al. (2006). "Evolution of alternative transcriptional circuits with identical logic." Nature **443**(7110): 415-420.

Tuch, B. B., D. J. Galgoczy, et al. (2008). "The evolution of combinatorial gene regulation in fungi." PLoS Biol **6**(2): e38.

Tuch, B. B., H. Li, et al. (2008). "Evolution of eukaryotic transcription circuits." Science **319**(5871): 1797-1799.

Vershon, A. K. and A. D. Johnson (1993). "A short, disordered protein region mediates interactions between the homeodomain of the yeast alpha 2 protein and the MCM1 protein." Cell **72**(1): 105-112.

Wach, A. (1996). "PCR-synthesis of marker cassettes with long flanking homology regions for gene disruptions in S. cerevisiae." Yeast **12**(3): 259-265.

Wagner, G. (2008). "Resurrecting the role of transcription factor change in developmental evolution." Evolution.

Wohlbach, D. J., D. A. Thompson, et al. (2009). "From elements to modules: regulatory evolution in Ascomycota fungi." Curr Opin Genet Dev: 1-8.

Wray, G. A. (2007). "The evolutionary significance of cis-regulatory mutations." Nat Rev Genet 8(3): 206-216.

Chapter 4:

Increased regulatory complexity as a consequence of functional interference between gene duplicates

Abstract:

Gene duplication can relax the level of constraint on gene evolution relative to the pre-duplication, ancestral gene. Yet, gene duplication can also introduce new constraints on gene evolution that did not exist in the pre-duplication state. An example is the constraint of functional interference between gene duplicates. A gene duplicate can exert a repressive effect on the function of the other duplicate if both proteins compete for binding an ancestral interaction partner, but only one of the duplicates retains the capacity to form a functional interaction with this partner. Here, we investigate the duplication in the hemiascomycete yeast lineage of a deeply conserved, multi-functional transcription regulator. Modifications to ancestral interactions in the evolutionary trajectory of both duplicates minimized the risk of functional interaction between the duplicates. One consequence of these modifications was an increase in the number of unique subunits within a regulatory complex. Thus, constraints on gene evolution introduced by gene duplication can act as an evolutionary ratchet, preventing the loss of increased regulatory complexity.

Introduction

Gene duplication provides an abundant resource of new genes across all domains of life (Zhang, Gaut et al. 2001). In gene duplication, new genes are conceived through stochastic mutational events that generate two (or more) copies of an ancestral gene. Two requirements must be fulfilled for both paralogs generated by a duplication event to persist over evolutionary time. First, it is necessary that both gene copies drift towards fixation across a species. Second, once fixed, it is necessary that the selection pressure for maintenance of both paralogs exceed the rate of degenerative mutations. How do new genes created by gene duplication make themselves indispensible to a species? For each maintained set of duplicated genes, the precise mutational path will be unique, but a common trend among characterized examples is the pattern of subfunctionalization (Lynch and Force 2000).

In subfunctionalization, the functions of the ancestral gene are distributed between the duplicated copies. Subfunctionalization may be adaptive, which can aid duplicated genes in sweeping to fixation across a population. In adaptive subfunctionalization, an ancestral, multi-functional gene cannot be optimized for all of its functions. The adaptive conflict between these functions is resolved in gene duplication as each paralog optimizes for a subset of the ancestral gene functions (this model of subfunctionalization is called ‘Escape from Adaptive Conflict’) (Innan and Kondrashov 2010). However, many subfunctionalizations may not require an adaptive explanation to account for their longevity in the genome (Lynch and Force 2000). In such instances, the partitioning of ancestral functions between two paralogs is the result of the differential degeneration of ancestral functions in each paralog. In this scenario, while conferring no selective advantage, both paralogs become necessary to complement the function of a single, ancestral gene. This type of subfunctionalization holds important implications for how we understand the evolution of genomic complexity (Lynch and Conery 2003). It should be noted that these two models of subfunctionalization are not mutually exclusive and may act synergistically to both promote the fixation of duplicated genes and their evolutionary longevity within the genome.

Here we combine ancestral gene reconstruction with computational and molecular biological techniques to investigate the gene duplication of a MADS-box transcription regulator on the phylogenetic branch leading to the model ascomycete yeast *S. cerevisiae*. The paralogs generated by the duplication of this ancestral MADS-box transcription regulator are known in *S. cerevisiae* as Mcm1 and Arg80 (Shore and Sharrocks 1995). We selected this specific case study because it allowed for investigation of some of the key principles influencing the evolutionary trajectory of long-term retained gene duplicates.

First, a rarely addressed constraint against the subfunctionalization of ancestral gene function is the challenge of functional interference between degenerating paralogs. Take the example of a transcription regulator. Transcription regulators form DNA-protein interactions at *cis*-regulatory sites throughout the genome and form (often many different) protein-protein interactions with other transcription regulators bound nearby on DNA. In subfunctionalization, each duplicate may lose some of these protein-protein interactions to other regulators. However, each duplicate will continue to compete for binding to the same DNA sequences, although they may not be competent to form all the interactions necessary to execute gene regulation at that site. In such a scenario, individuals that contain both duplicates will be at disadvantage to the pre-duplication ancestor. The duplication of the ancestral MADS-box regulator transcription regulator makes an ideal case study to explore the importance of functional interference between paralogs. The ancestral MADS-box protein regulated on the order ~4% of the genes in the genome and formed protein-protein interactions with a variety of other transcription regulators (Tuch, Galgoczy et al. 2008).

Second, gene duplication plays an important role in the creation of new subunits within molecular complexes (Finnigan, Hanson-Smith et al. 2012). As the components of molecular complexes have been characterized, these complexes often contain homologous proteins forming interactions with one another (Lander, Estrin et al. 2012) (Angel and Karin 1991). The relationship between gene duplication and the structure of molecular complexes raises the question: are new subunits incorporated into these complexes for adaptive reasons or are they being retained through the neutral mechanisms, such as reciprocal loss of function, that often tend to fix gene duplicates in the genome (Finnigan, Hanson-Smith et al. 2012)? In our case study, the ancestral MADS-box transcription regulator participated in a host of regulatory complexes: binding DNA as a homodimer and forming additional protein-protein contacts to other DNA bound transcription regulators (Tuch, Galgoczy et al. 2008). Following duplication, the paralogs maintain this ancestral dimerization surface but participate now in both homo- and heterodimerization between paralogs (Messenguy and Dubois 1993). Thus, a molecular dissection of the events that biased the paralogs towards homo- versus heterodimerization creates may provide insight into the evolutionary circumstances leading to either the retention of both or a single paralog within a molecular complex.

Our results demonstrate that the last common ancestor of the two MADS-box paralogs (Arg80 and Mcm1) is an effective complement for both paralogs in *S. cerevisiae*, consistent with non-adaptive evolutionary mechanisms being the primary explanation for the retention of both duplicates. The ancestral MADS-box protein formed interactions with multiple cofactor transcription regulators using a shared set of amino-acid residues (Mead, Bruning et al. 2002). We find that mutations to these interacting residues in the paralogs simultaneously strongly weakened the interaction with certain cofactors in the daughter paralogs and may have subtly optimized the interaction with certain cofactors. The evolutionary trajectory of these gene duplicates appears to have been shaped by the risks of functional interference between degenerating paralogs. We find that modifications to protein-protein interactions and the weakening of the DNA-binding affinity of Arg80 minimized the risk of functional interference. As a result of its reduced DNA binding affinity, Arg80 tends to form a heterodimer complex with Mcm1 to stabilize its DNA binding (instead of the unstable Arg80 homodimer). Thus, the formation of a heterodimeric complex between paralogs in post-duplication species appears unlikely to have been an adaptation that enhanced gene regulation driven by selection. Rather, this increase in regulatory complexity seems to be a consequence of constraints on the evolution of gene duplicates, such as functional interference between paralogs. Given the frequency of multiple paralogs interacting within molecular complexes, the evolutionary trajectory described here may act as model to understand how the incorporation of multiple paralogs into molecular complexes may be maintained by evolutionary pressures other than selection acting to increase the fitness of the complex over what was possible in the pre-duplication state.

Results

Duplication of a multifunctional MADS-box transcription regulator

For three reasons, the ancestral MADS-box transcription regulator in the hemiascomycete yeast lineage would seem a likely candidate for the optimization of individual gene functions following gene duplication. First, the ancestral MADS-box regulator formed interactions with a number of other transcription regulators and participated in regulation of a substantial fraction of the genes in the genome (Tuch, Galgoczy et al. 2008). These protein-protein contacts to other transcription regulators (combinatorial interactions) target transcription regulators to different groups of functionally related genes, allowing for these functional gene sets to be coordinately regulated. For the MADS-box transcription regulator in the yeast (Mcm1 and Arg80), these combinatorial interactions include an interaction with the transcription regulator α1 that targets Mcm1 to a group of genes required for mating called the α-specific genes and an interaction with the transcription regulator Arg81 that targets Arg80 to the arginine metabolic genes (ARG genes) (Amar, Messenguy et al. 2000) (Bender and Sprague 1987) (Figure 1A). The interactions of α1 and Arg81 with the MADS-box transcription regulator are an ancient trait (Tuch, Galgoczy et al. 2008). Thus, at the time of the duplication of the ancestral MADS-box transcription regulator, this protein balanced the interaction with many different transcription regulators like Arg81 and α1 (Figure 1A).

Second, we know from extensive molecular biological work on Arg80 and Mcm1 in *S. cerevisiae* that some of these combinatorial interactions are formed using overlapping sets of residues within the MADS-box domain (Mead, Bruning et al. 2002)

(Jamai, Dubois et al. 2002). For instance, residues at homologous positions within Mcm1 and Arg80 are critical for the interaction of Mcm1 with α1 and Arg80 with Arg81. This implies that prior to gene duplication, Arg81 and α1 would have used many of the same residues on the ancestral MADS-box transcription regulator to form their respective interactions.

The tandem duplication of the ancestral MADS-box gene occurred in the last common ancestor of *Z. rouxii/S. cerevisiae* (Figure 1B). Following duplication, the two paralogs experienced unequal rates of molecular divergence. The Arg80 paralog extensively diverged from the ancestral MADS-box domain sequence. Ascomycete fungal evolution has been approximated as dating back 400 mya to 1.8 bya (Taylor and Berbee 2006). Over these epochs, the 86 amino acid MADS-box domain (the region of the transcription regulator responsible for forming both DNA and protein contacts) experienced just 4 amino-acid changes (Figure S1). Following duplication, the common ancestor of all Arg80 paralogs differed from the pre-duplication state by 17 amino-acid changes within the MADS-box domain (Table S1). In contrast, the common ancestor of all Mcm1 paralogs diverged at just two amino acids from the pre-duplication state within the MADS-box domain. It should be noted that these two changes still represented a significant divergence relative to the slow rate of evolution in the pre-duplication MADS-box protein.

This difference in the rate of molecular divergence between paralogs translated to differences in function. The Mcm1 paralog is an essential gene that has retained many of the ancestral functions of the MADS-box protein, including the regulation of cell cycle genes, mating genes (including the α-specific genes), and the ARG genes (Shore and

Sharrocks 1995). In contrast, Arg80 is a non-essential gene that specialized to exclusively regulate the ARG genes.

Yet, the two paralogs share many similarities as a consequence of their descent from a common ancestor. They both participate in the regulation of the ARG genes in *S. cerevisiae* as a heterodimer of paralogs at the ARG genes. Second, as is the case for many paralogs, they share closely related DNA-binding specificities (Boonchird, Messenguy et al. 1991) (Hayes, Sengupta et al. 1988). The impact of this shared DNA-binding specificity can be observed by over-expressing Arg80 (Figure 1C-D). Arg80 does not participate in the regulation of the α-specific genes; this is exclusively a function of Mcm1. Yet, over-expression of Arg80 actually diminishes the level of α-specific gene expression. The repressive effect of Arg80 on α-specific gene expression is a reflection of Arg80 outcompeting Mcm1 for binding to *cis*-regulatory sequences at this group of genes. Arg80 can do this because it retains a similar DNA-binding specificity and the increased Arg80 protein levels favor binding of the incorrect paralog at these regulatory sites.

The molecular biology and evolution of the Arg80 and Mcm1 paralogs raised a number of questions to us. Why are both paralogs retained when they have overlapping functions? Does the incorporation of both paralogs into the ARG gene regulatory reflect some adaptation over the ancestral regulatory mode? Furthermore, the shared properties of Arg80 and Mcm1 (such as their DNA-binding specificity) can give rise to functional interference between the paralogs. Has the evolutionary process minimized this risk and if so, how?

Efficient complementation of Arg80/Mcm1 by the pre-duplication ancestor

To address these questions, we synthesized ancestral versions of the MADS-box genes. Ancestral gene reconstruction is an approach that has proven useful for testing evolutionary predictions (Thornton 2004). The strategy depends on the accurate protein alignment of the ortholog group of interest, followed by the calculation of amino acid probabilities at each position within the ancestral protein using a species or gene tree as a guide. We synthesized ancestral MADS-box domains from the three nodes relevant to the gene duplication—the last common ancestor of all Mcm1 paralogs (AncMcm1), the last common ancestor of all Arg80 paralogs (AncArg80), and the last pre-duplication ancestor (AncMADS). We focused exclusively on reconstructing ancestral MADS-box domains because this region is sufficient for forming interactions with DNA, as well as the protein-protein interactions with α1 and Arg81 (Qiu, Dubois et al. 1990) (Mead, Bruning et al. 2002).

The capacity of a pre-duplication ancestral gene to complement a deletion of the modern paralogs in a post-duplication species will depend on the evolutionary trajectories of the gene duplicates. For instance, if the paralogs have undergone adaptation relative to the pre-duplication gene (such as in neofunctionalization or Escape from Adaptive Conflict), then the pre-duplication ancestor will fail to complement the modern paralogs. In non-adaptive models (such as canonical subfunctionalization), the pre-duplication ancestor may act as an efficient complement for the modern paralogs.

In our case, AncMADS complemented a deletion of both the *S. cerevisiae* Arg80 and Mcm1 paralogs (Figure 2A-D, Figure S2). Most Mcm1 regulated genes showed no statistical difference when AncMADS replaced both paralogs (Figure S2). The pre-duplication ancestor was not a perfect complement. Two groups of MADS-box regulated genes (the α-specific genes and the ARG genes) had a diminished dynamic range in the AncMADS strain relative to wild type (Figure 2A-C). The dynamic range of ARG gene expression was also weaker than a strain where AncArg80 replaced the endogenous Arg80. However, the differences are small. The changes in ARG gene regulation when AncMADS replaced Mcm1 and Arg80 did not translate to a phenotypic effect (Figure 2D). Furthermore, the MADS-box proteins are components within gene regulatory networks that have continued to evolve for an estimated 80 to 150 mya since the pre-duplication state (Taylor and Berbee 2006). In this light, that the small difference between complement by AncMADS and the wild type is likely the result of evolution in the transcriptoinal networks around MADS-box proteins.

These results indicate that the post-duplication evolutionary trajectory of Arg80 and Mcm1 can be best described as a non-adaptive subfunctionalization. Yet, these results also raise a new question— why were both duplicates maintained if the pre-duplication gene effectively performs the functions of both duplicates?

The reciprocal loss of ancestral interactions

S. cerevisiae Mcm1 and the ancestral MADS-box protein formed many combinatorial interactions to other transcription regulators using overlapping residues in the MADS-box domain. For instance, the interactions between Arg80/Arg81 and Mcm1/α1 depend heavily on a homologous set of residues within the MADS-box domain that form a binding pocket (Mead, Bruning et al. 2002) (Jamai, Dubois et al. 2002) (Figure 3A-C). The evolutionary history of three crucial residues within this pocket structure (positions 110, 118, and 119) caught our attention. One of these residues was identified as necessary for the interaction of Mcm1 with α1 (position 110) and all three have been identified as critical for the interaction between Arg80 and Arg81. Position 110 has diverged in Mcm1 ($Y_{110}F$) with a removal of a hydroxyl group from tyrosine to form phenylalanine, but remains in the ancestral state in Arg80 (Figure 3D). In contrast, positions 118/119 have diverged only in Arg80 ($TQ_{118/119}AN$) causing a substantial change in the architecture of the binding pocket in that paralog.

The capacity of the AncMADS gene to regulate the ARG genes is eliminated by introducing the $Y_{110}F$ mutation into this protein (Figure 3E-F). The effect of this mutation is specific to the ARG genes as Mcm1 regulated genes as most Mcm1 regulated transcripts are unaffected by this mutation (Figure S1). The exception is the α-specific genes, where the $Y_{110}F$ mutation improves the dynamic range of gene regulation to wild-type levels. This result strongly supports that the interaction between the AncMADS protein and Arg81 has been compromised since the effect is specific to the ARG genes and prior work has identified this position as a core component of the interaction between Arg80 and Arg81.

We made the prediction that if the $Y_{110}F$ mutation weakened the interaction between Mcm1 and Arg81 in the post-duplication ancestor, then this mutation should also weaken the interaction with an Arg81 protein from a pre-duplication species. *K. lactis* is a hemiascomycete yeast species that branches before the MADS-box gene duplication. The *K. lactis* Arg81 gene complemented a deletion of the *S. cerevisiae* Arg81 (Figure 3F). However, this complementation depended on the presence of Arg80 (which contains the Y_{110} residue). Thus, through the loss of an ancestral interaction by the Mcm1 paralog, Arg80 became essential to maintain regulation of the ARG genes.

Yet, when Mcm1 lost the capacity to interact with Arg81, why was this paralog not eliminated from the genome through additional degenerative mutations? The MADS-box domain of the Arg80 paralog significantly diverged from the sequence of the pre-duplication ancestor and two of these mutations ($TQ_{118/119}AN$) map to the same binding pocket as the Y_{110} residue (Figure 3A-D). Introducing the $TQ_{118/119}AN$ into the AncMADS mutations eliminated its capacity to interact with $\alpha1$ and regulate the α-specific genes (Figure 3G). Thus, whereas mutations in Mcm1 weakened its interaction with Arg81, mutations in Arg80 weakened different ancestral interactions, such as the interaction with $\alpha1$. A trivial possibility was that the $TQ_{118/119}AN$ mutations destabilized the AncMADS protein, resulting in a partially non-functional protein that was compromised in the regulation of all MADS-protein regulated genes. However, this is not the case, many Mcm1 regulated genes are unaffected by the AncMADS $TQ_{118/119}AN$ mutant and this mutant actually improves the dynamic range of ARG gene regulation over the non-mutant AncMADS gene sequence (Figure S1). We also tested whether these mutations impaired the capacity of the AncMADS sequence to interact with a pre-duplication $\alpha1$ protein. The *K. lactis* $\alpha1$ gene can complement *S. cerevisiae* $\alpha1$ (Baker, Tuch et al. 2011). Again, the $TQ_{118/119}AN$ mutations impaired the capacity of AncMADS to regulate the α-specific genes (Figure 3G). This finding further supports that mutations specific to the Arg80 paralog, (such as $TQ_{118/119}AN$) impaired the capacity of this protein to form ancestral interactions with other transcription regulators, such as $\alpha1$.

In total, the retention of both MADS-box paralogs over evolutionary time can be attributed to the reciprocal loss of ancestral interactions in both paralogs. Although small in magnitude, we also observed that introducing derived residues into the pre-duplication MADS-box ancestor improved its capacity to regulate subsets of the MADS-box regulated genes ($Y_{110}F$ improved α-specific gene regulation and $TQ_{118/119}AN$ improved ARG gene regulation). Whether a result of adaptation or drift, these modifications to specific cofactor interactions in each paralog have important consequences for transcriptional regulatory networks. One of these consequences, discussed further below, is that how they minimized the risk of functional interference between paralogs.

The Reduced DNA-binding affinity of Arg80

With Mcm1 having lost its capacity to interact with Arg81, one might predict that Mcm1 would not retain a role in regulating the ARG in post duplication species. However, unique to the ARG genes, Mcm1 and Arg80 form a heterodimer of paralogs, with Arg80 interacting directly with Arg81. Why is this gene set regulated by a MADS-box heterodimer? Has the incorporation of both paralogs into the ARG gene regulatory complex allowed for some adaptations over the ancestral complex?

The answer to these questions rests in the relative DNA-binding affinities of the paralogs. We assayed the half-life on a *S. cerevisiae* ARG gene *cis*-regulatory sequence (*ARG3*) of the different ancestral MADS-box proteins (Figure 4). The pre-duplcation AncMADS and AncMcm1 proteins had identical half-lives on DNA, but the half life of the AncArg80 paralog on DNA was substantially decreased. Thus, with no other factors influencing DNA-binding but *cis*-regulatory site and the MADS-box proteins, Mcm1 would outcompete Arg80 for binding to the ARG genes. That this is not the case in post-duplication reflects the stabilizing effect of Arg81 on Arg80 binding. Yet, Arg81 interacts with a single side of the MADS-box dimer and will not exert a stabilizing effect on the subunit it does not directly contact. Interaction energies will favor that this second subunit is the Mcm1 paralog because of its longer half-life on DNA.

In this light, the incorporation of both paralogs into the ARG gene regulatory complex does not appear as an innovation that provided new regulatory potential to the complex. Rather, it may be an unintended consequence of weakened ancestral interactions (the reduced DNA-binding affinity of the Arg80 paralog). A decrease in Arg80 DNA-binding affinity also further tipped the balance in favor of Mcm1 binding at gene sets exclusively regulated by the Mcm1 paralogs (such as the α-specific genes). This event again minimized the risk of functional interference between the post-duplication MADS-box proteins.

Discussion

The *S. cerevisiae* Arg80 and Mcm1 transcription regulators descend from a tandem gene duplication event in the last common ancestor of *S. cerevisiae*/*Z. rouxii* (Figure 1B). Through a combination of ancestral gene reconstruction, computational, and molecular biological experiments, we describe the molecular events that underlie the evolutionary ratchet preventing the loss of either paralog. The ancestral MADS-box protein efficiently complemented both paralogs in *S. cerevisiae* (Figure 2A-D). Thus, if this gene duplication increased fitness over pre-duplication ancestors, the effect must have been minor. The evolutionary path of the gene duplicates also revealed evidence of the constraint functional interference placed on the subfunctionalization of ancestral gene function between these degenerating paralogs (Figure 1C-D, 3E-G, 4). We found that the risk of functional interference between the Arg80 and Mcm1 paralog was minimized through the emergence of a reduced DNA-binding affinity in Arg80 and modifications in ancestral protein-protein interactions). A consequence of the reduced DNA-binding affinity of Arg80 was the creation of a preference for an Arg80/Mcm1 heterodimer at the ARG genes (Figure 1A). Therefore, the involvement of both paralogs in regulation of this gene is unlikely to reflect a more optimized regulatory state, but instead may have been the consequence of constraints on Arg80 gene evolution arising from functional interference between paralogs.

In this discussion we first address the molecular events that allowed for the long-term retention of both paralogs, with an emphasis on why adaptation may have played an insignificant role in these events. Second, we elaborate on the general importance of functional interference for understanding the evolution trajectory of paralogs during subfunctionalization. Finally, we discuss the evolution of molecular complexes and the molecular events that can incorporate new subunits into complexes independent of adaptation.

The work of one divided in two

The success of gene duplication in expanding the number of coding sequences in the genome can obscure the fact that gene duplicates face many challenges in reaching fixation and to their retention over evolutionary time. For long-term retention, the selection pressure to maintain both gene duplicates must be greater than the rate of degenerative mutations. We find that selection pressure to maintain both Arg80 and Mcm1 is rooted in the reciprocal loss of interactions by both paralogs. In each paralog, ancestral interactions with one (or more) cofactor transcription regulators have been compromised. For Mcm1, a subtle mutation ($Y_{110}F$) substantially weakened its for affinity for the transcription regulator Arg81 at the ARG genes (Figure 3E-F). Within the same binding pocket as the $Y_{110}F$ Mcm1, two residues in Arg80 ($TQ_{118/119}AN$) weakened the interaction its interaction with the transcription regulator α1 at mating-type regulated genes. Thus, to maintain ancestral functions of the pre-duplication gene, the loss of ancestral interactions acts as an evolutionary ratchet that blocks the loss of either paralog.

From this perspective, there is no rule that the post-duplication state must be more fit than the pre-duplication state. Yet, for good reasons, gene duplication has been linked to opportunities for adaptation. One model that links gene duplication to adaptation is neofunctionalization (Innan and Kondrashov 2010). In neofunctionalization, a new function evolves in one of the two duplicates. This novelty can arise because one of the paralogs can explore evolutionary space while the other duplicate retains the ancestral function. It has been noted that many gene duplicates show unequal rates of evolution between paralogs and on occasion, this fact has been cited as evidence for neofunctionalization (Byrne and Wolfe 2007). Arg80 and Mcm1 also follow this pattern of unequal evolutionary rates following gene duplication (Figure 1B). Yet, no new functions have been discovered in either paralog relative to the ancestral MADS-box protein (Tuch, Galgoczy et al. 2008). More likely, the unequal evolutionary rates between Arg80 and Mcm1 are a reflection of the unequal distribution of ancestral functions between Mcm1 and Arg80.

A second model linking adaptation to gene duplication is ‘Escape from Adaptive Conflict’. In escape from adaptive conflict, the multiple function of an ancestral gene in conflict and in gene duplication, each paralog can be optimized for a subset of those ancestral functions. The duplication of Arg80 and Mcm1 shares much in common with this model. First, the ancestral MADS-box protein was multifunctional and these different functions were critical to survival in the wild. Second, the duplication allowed for mutations that would have been strongly selected against in the ancestor to take place in the duplicated paralogs (such as the mutation $Y_{110}F$ in Mcm1 and the mutation of $TQ_{118/119}AN$ in Arg80). Finally, these changes had a measurable positive effect on the effectiveness of each duplicate. In Mcm1, $Y_{110}F$ increased the affinity of Mcm1 for its cofactor protein $\alpha1$ at the mating-type regulated genes and in Arg80, the $TQ_{118/119}AN$ mutation increased the affinity of that paralog for its cofactor transcriptional regulator Arg81 at the ARG genes (Figure S1).

Yet, do these elements constitute evidence for an escape from adaptive conflict? Two observations raise doubts about whether ‘Escape from Adaptive Conflict’ describes the evolutionary trajectory of the post duplication MADS-box genes. First, the effect on gene expression produced by weakening interactions (Arg80 with $\alpha1$, Mcm1 with Arg81) is far greater than the impact of lineage specific mutations appearing to optimize interactions (Mcm1 with $\alpha1$, Arg80 with Arg81). Second, does the increase in the strength of a cofactor interaction constitute ‘an optimization’ over the ancestral gene? The total free energy for a transcriptional regulatory complex to occupy a given *cis*-regulatory sequence is the sum of all the protein-protein interactions and DNA-protein interactions involved. Thus, the weakening of DNA-binding interactions can offset the strengthening of protein-protein interactions. Exactly such an event appears to have occurred for Arg80, where a strengthened interaction with Arg81 may have offset the loss in DNA-binding affinity by this paralog. Thus, what appears like the 'optimization' of Arg80 and Mcm1 may actually be the process of redistributing interaction energies within transcriptional regulatory complexes. In the section below, we will discuss why this redistribution may have taken place.

In conclusion, we do not see a major role for adaptation in the retention of both duplicates following the duplication of the ancestral MADS-box protein. The single most important factor influencing retention has likely been the reciprocal loss of interactions by both paralogs. The effect of optimization of the two duplicates is at best minimal and may not be an optimization at all, but rather a reflection of propensity of transcriptional regulatory complexes and their binding sites to redistribute interaction energies over evolutionary time.

The constraints introduced by gene duplication

Escape from adaptive conflict describes a path for subfunctionalization following gene duplication to increase fitness. The opposite situation would be one where subfunctionalization of an ancestral function between paralogs decreases fitness. For the duplication of proteins forming many interactions (such as transcription regulators), this latter scenario is an intrinsic risk in subfunctionalization. A gene duplicate undergoing subfunctionalization has the potential to interfere with the function of its sister paralog.

In our case study, Arg80 evolved mutations that broke the ancestral interaction with $\alpha 1$ and its capacity to activate expression of mating-type specific genes. The loss of this protein-protein interaction between Arg80 and α1 tilted the equilibrium for binding to the α-specific gene *cis*-regulatory sequences towards the paralog Mcm1 (which retains the capacity to form an interaction with α1). Yet, even lacking this interaction, Arg80 will compete with Mcm1 and win some fraction of the binding events to α-specific genes *cis*-regulatory sequences because the two paralogs share the same DNA-binding specificities (due to their descent from a common ancestor). The impact of this competition can be detected by over-expressing Arg80. Although Arg80 does not regulate α-specific genes in *S. cerevisiae*, over-expression of Arg80 drives down the expression of α-specific genes by 4-fold.

Along with the modification of protein-protein interactions, the risk of functional interference between Arg80 and Mcm1 has been minimized through a decrease in the DNA-binding affinity of Arg80. Taken together, the weakening of ancestral protein-protein interactions, slight strengthening of different ancestral protein-protein interactions, and the weakening of DNA-binding affinity in one paralog each act to reinforce the block to functional interference between paralogs. We cannot known for certain whether selection or neutral drift led to a molecular event such as the weakening of Arg80 DNA-binding affinity. However, it is clear that in addition to removing constraints on gene evolution, gene duplication can introduce new constraints that did not exist in the pre-duplication state. To minimize the risk of functional interference, gene duplicates can follow evolutionary trajectories that modify ancestral interactions to limit competition between duplicates. Despite the risks of functional interference, species where natural selection acts inefficiently (e.g.- species in small population sizes) may be forced to tolerate functional interference and duplicates that reach fixation may decrease fitness relative to the pre-duplication state.

Incorporation of new subunits into regulatory complexes

Sophisticated molecular complexes are responsible for much of the biology of the cell. When we ask questions about the evolution of these complexes, a challenge is that removal of a single subunit from a complex can often collapse function of the entire unit. Yet, quite often these molecular complexes contain gene duplicates. Thus, subfunctionalization can act as an explanation for the incorporation of new subunits into complexes (and why both paralogous subunit become essential following the partitioning of ancestral functions). This also suggests that the incorporation of new subunits to molecular complexes need not be driven by adaptation.

The ancestral MADS-box transcription regulator bound DNA as a homodimer and formed protein-prortein interactions with a number of cofactor transcription regulators bound to adjacent DNA sequences. At the ARG genes, this cofactor transcription regulator interacting with the ancestral MADS-box protein was Arg81. Post duplication, the most common regulatory architecture at this gene set is a heterodimer of Arg80 and Mcm1 with Arg80 supporting the protein-protein interaction to Arg81. With gene duplication, both paralogs have been incorporated into this regulatory complex. This is in contrast to the circumstances at other gene sets regulated by the ancestral MADS-box regulator, such as the mating-genes where a Mcm1 homodimer regulates these gene sets.

The molecular explanation for the formation of a heterodimer of paralogs within the ARG gene regulatory complex rests in the decreased DNA-binding for the Arg80 paralog. The Arg80 homodimer is unstable and therefore, interaction energies favor the more stable complex of Mcm1/Arg80/Arg81 at the ARG gene *cis*-regulatory sequences than an Arg80 homodimer/Arg81 complex.

The evolutionary trajectory to the incorporation of both paralogs into the ARG gene regulatory complex is marked by loss of function events (loss of the an interaction between the Mcm1 paralog and Arg81, weakening of Arg80 DNA-binding). Thus, it seems unlikely that both duplicates were incorporated into this complex as a means to optimize regulation to levels unachievable in the pre-duplication state. If this process was not driven by adaptation for improved fitness of the regulatory complex, then why did the molecular events that led to the incorporation of both duplicates into this regulatory complex occur?

We demonstrated that conditions giving rise to functional interference between these paralogs have a measurable cost on regulation (Figure 1D). The molecular events that led to this increase in regulatory complexity are a subset of the molecular events that also reduced the functional interference between the Arg80 and Mcm1 paralogs. From these observations, we concluded that the cost of functional interference has been a key constraint (if not the exclusive constraint) maintaining the heterodimeric state. Evolutionary paths that would stabilize an Arg80 homodimer, such as increasing Arg80 DNA-binding affinity, will also increase the competition between Mcm1 and Arg80 at gene-sets where Arg80 is non-functional. It may not be necessary to invoke natural selection as the guiding force that drove the events that reduced functional interference; neutral drift may have chanced upon these molecular solutions. Instead, it is cost of functional interference that creates an evolutionary ratchet preventing the loss of the heterodimer of paralogs within the ARG gene regulatory complex.

Many transcription regulatory complexes contain heterodimers formed by gene duplicates (Angel and Karin 1991) (Leid, Kastner et al. 1992). Our work shows how the evolutionary forces that maintain the presence of both duplicates within a regulatory complex may be unrelated to the direct regulatory function of the complex and may be difficult to rationalize without applying an evolutionary framework.

Concluding Remarks

The gene duplication of the ancestral MADS-box regulator in the hemiascomycete yeast highlights how the structures of modern transcription regulatory complexes are shaped by their evolutionary history. Constraints that may not be readily apparent in modern species (such as the functional interference between degenerating paralogs) may drive evolutionary events that remodel the structure of these regulatory complexes. In this instance, functional interference between the duplicates of an ancestral multifunctional transcription regulator stabilized a series of molecular events that led to the incorporation of both duplicates into an ancestral regulatory complex. Thus, it appears that gene duplication necessitated an increase in regulatory complex, as opposed to making such an event possible. We hope that these insights can provide a general model to understand the prevalence of heterodimers formed by gene duplicates.

Materials & Methods

Ancestral Reconstruction of MADS-box domains

Orthologs of S. cerevisiae Mcm1 and Arg80 were defined by a tBLASTN search of the 35 additional hemiascomyete genomes available at the NCBI website, as well as a BLASTp search of the collection of yeast genomes available only on the Yeast Gene Order Browser webpage (Gordon, Armisén et al. 2011) (http://wolfe.gen.tcd.ie/ygob/). To eliminate false positives, hit sequences were reverse BLASTed against the *S. cerevisiae* genome. A protein alignment was built from this collection of sequences through submission to PRANK (Löytynoja and Goldman 2005). A gene tree for these sequences was generated by PhyML (Figure 1B) (Guindon and Gascuel 2003). Specific parameters used within PhyML that differ from default settings included: use of the JTT model for amino-acid substitutions, the calculation of approximate likelihood ratios (yes/aL), and the tree topology model selection setting was set to choose a best fit between NNI and SPR models.

A second gene tree was generated from the same protein sequence alignment and the same settings, except without approximate likelihood ratios, for submission to Lazarus for the reconstruction of ancestral sequences (Hanson-Smith, Kolaczkowski et al. 2010). Beyond the default parameters, we chose to set the amino-acid substitution model to JTT, ancestral gaps sequence were allowed, and the *Y. lipolytica* sequence was set as the out-group. Nearly all amino-acid positions within the MADS-box domain were reconstructed to certainties of greater than 90% at our branchpoints of interest.

RNA isolation & quantification

Strains were grown in YEPD to OD_{600} = 0.8 and then centrifuged at 4000 rpm for 5 minutes. The supernatant removed and pellets frozen in liquid nitrogen. RNA was isolated as previously described (Mitrovich, Tuch et al. 2007) with all volumes scaled appropriately. Total RNA was quantified by OD_{260} and its purity assessed using OD_{260} /OD_{230} and OD_{280} /OD_{260} ratios. NanoString quantification of transcript abundance was performed by NanoString Core facility in Seattle, Washington, USA.

Ornithine growth assay

Cells were grown overnight in YEPD and then inoculated at an OD_{600} of .1 into a minimal ornithine growth media, which included only glucose, yeast nitrogen bases (without ammonium), and ornithine as the sole nitrogen source for amino-acid production.

Gel Shift experiments

Ancestral MADS-box proteins were expressed in *E. coli* BL21 (DE3) cells from the pET26b plasmid. 100 ml of cells were induced at an OD_{600} of .6 with 1 mM IPTG and incubated overnight at 16 °C and 300 rpm. Cells were lysed, proteins purified, and epitope tags removed from the recombinant protein as previously described (Lohse, Zordan et al. 2010). *S. cerevisaie ARG3 cis*-regulatory sequence oligonucleotide probes were labeled with P^{32} γ-ATP using T4 PNK. Binding conditions were 50 mM Tris [pH = 8], 100 mM NaCl, 10% Glycerol, 5 mM $MgCl_2$, 5mM β-mercapoethanol, 50μg/mL Poly(dI-dC) (limits non-specific protein:DNA-binding), and 1.2 μM labeled oligonucleotide. Labeled DNA was incubated in the presence of various concentrations of ancestral MADS-box proteins for 30 minutes. After 30 minutes, 120 μM unlabeled oligonucleotide was added to the reaction.

Strain Construction

A complete list of all strains and primers used in this study can be found in supplemental tables.

Figures

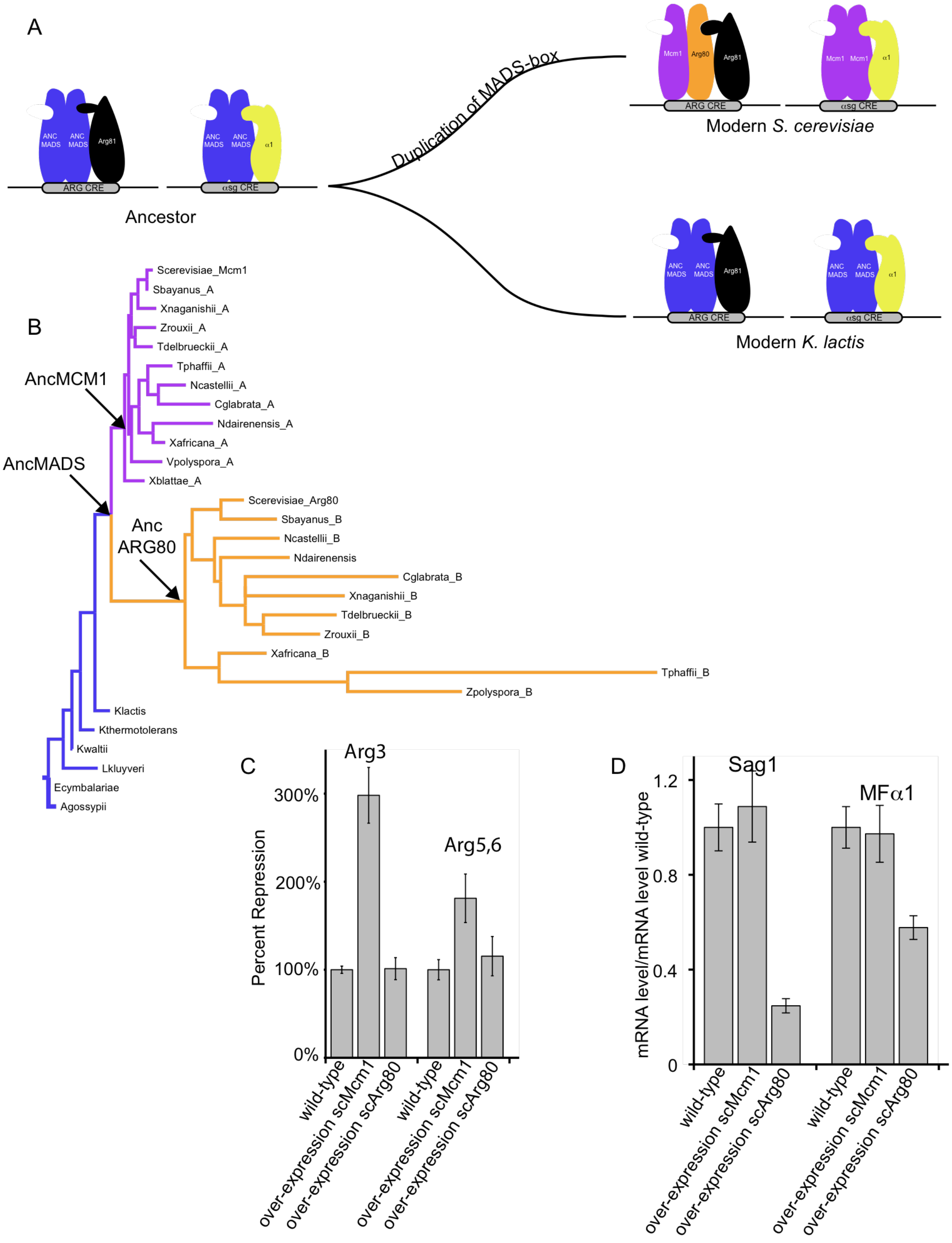

A
ANC MADS
ANC MADS
Arg81
ARG CRE
ANC MADS
ANC MADS
α1
αsg CRE
Ancestor
Duplication of MADS-box
Mcm1
Arg80
Arg81
ARG CRE
Mcm1
Mcm1
α1
αsg CRE
Modern S. cerevisiae
ANC MADS
ANC MADS
Arg81
ARG CRE
ANC MADS
ANC MADS
α1
αsg CRE
Modern K. lactis
B
AncMCM1
AncMADS
Anc ARG80
Scerevisiae_Mcm1
Sbayanus_A
Xnaganishii_A
Zrouxii_A
Tdelbrueckii_A
Tphaffii_A
Ncastellii_A
Cglabrata_A
Ndairenensis_A
Xafricana_A
Vpolyspora_A
Xblattae_A
Scerevisiae_Arg80
Sbayanus_B
Ncastellii_B
Ndairenensis
Cglabrata_B
Xnaganishii_B
Tdelbrueckii_B
Zrouxii_B
Xafricana_B
Tphaffii_B
Zpolyspora_B
Klactis
Kthermotolerans
Kwaltii
Lkluyveri
Ecymbalariae
Agossypii
C
Arg3
Arg5,6
Percent Repression
300%
200%
100%
0%
wild-type
over-expression scMcm1
over-expression scArg80
wild-type
over-expression scMcm1
over-expression scArg80
D
Sag1
MFα1
mRNA level/mRNA level wild-type
1.2
0.8
0.4
0
wild-type
over-expression scMcm1
over-expression scArg80
wild-type
over-expression scMcm1
over-expression scArg80

Figure 1 – The duplication and divergence of a multifunctional transcription regulator

(A) Model depicting the functional evolution of the MADS-box duplicates at two ancestrally –regulated gene-sets (the ARG and α-specific genes). **(B)** Gene tree of hemiascomycete MADS-box domain proteins. Blue branches are species that branch pre-duplication, orange branches are orthologs that cluster with *S. cerevisiae* Arg80, and purple branches are ortholgos that cluster with *S. cerevisiae* Mcm1. **(C-D)** Over-expression of *S. cerevisiae* MADS-box proteins. Expression data was collected using NanoString. Median and Standard Error were determined using an N of 3 replicates. **(C)** Impact of over-expressing Arg80 on repression of genes repressed by ARG gene regulatory complex. **(D)** Impact of over-expressing Arg80 on activation of the α-specific genes.

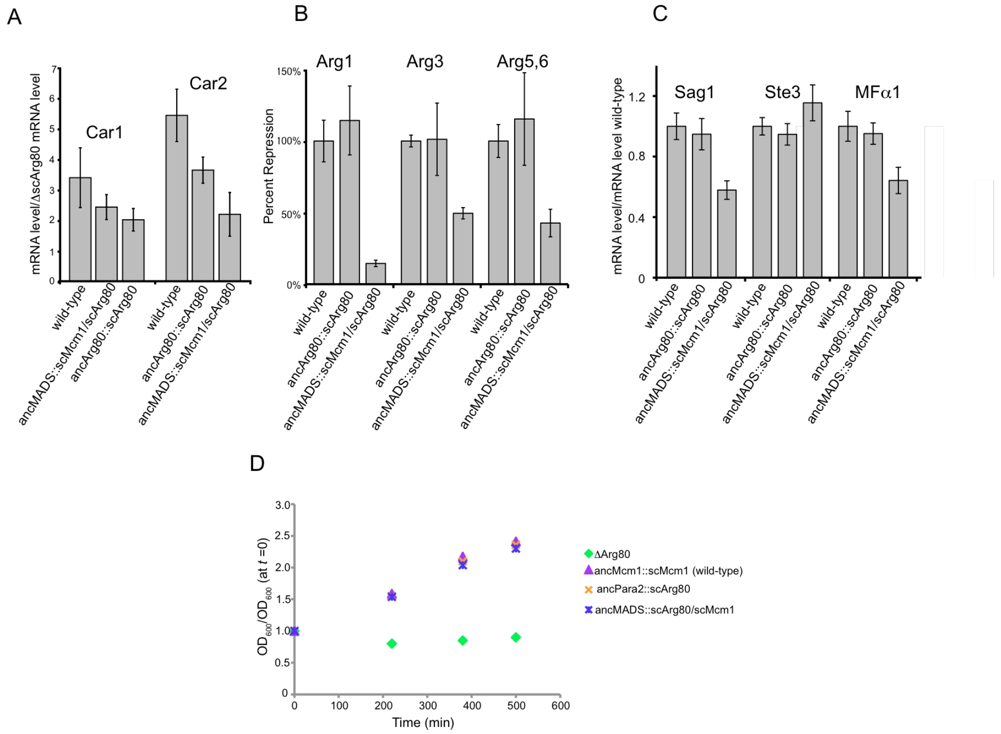


Figure 2 – The pre-duplication ancestor complemented both paralogs in *S. cerevisiae*

(A-C) The impact of replacing both endogenous *S. cerevisiae* MADS-box proteins with the pre-duplication MADS-box ancestor on gene expression. Gene expression quantified using NanoString. **(A)** Genes activated by the ARG gene regulatory complex. **(B)** Genes repressed by the ARG gene regulatory complex. **(C)** α-specific genes. **(D)** Growth of ancestral MADS-box gene strains using ornithine as a sole nitrogen source. Ornithine is converted into arginine and then modified to produce the other essential amino acids. In the absence of a functional ARG gene regulatory complex, strains cannot utilize ornithine as a nitrogen source. Median and Standard Error were determined using an N of 3 replicates.

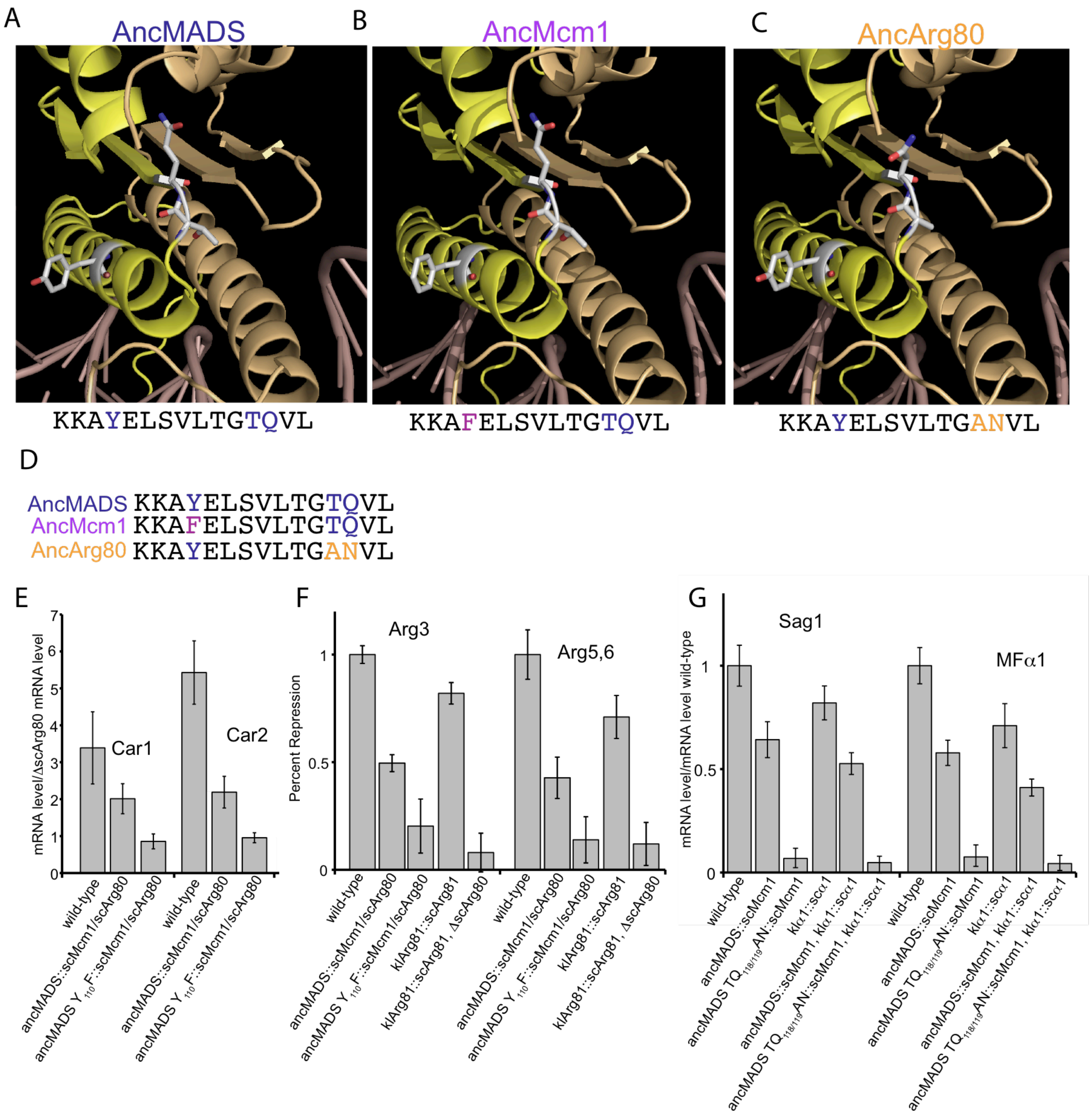


Figure 3 – Duplicates were maintained through the reciprocal loss of ancestral interactions

(A-C) The Mcm1 crystal structure ((Tan and Richmond 1998) with interacting side chains in grey. **(A)** AncMADS **(B)** AncMcm1 **(C)** AncArg80. **(D)** Alignment of the AncMcm1, AncArg80, and AncMADS with interacting side chain residues in bold. The impact of mutations on the complement of endogenous *S. cerevisiae* MADS-box proteins by ancestral proteins. Gene expression quantified using nanostring (**E**) Genes activated by the ARG gene regulatory complex. **(F)** Genes repressed by the ARG gene regulatory complex. **(G)** α-specific genes. Median and Standard Error were determined using an N of 3 replicates.

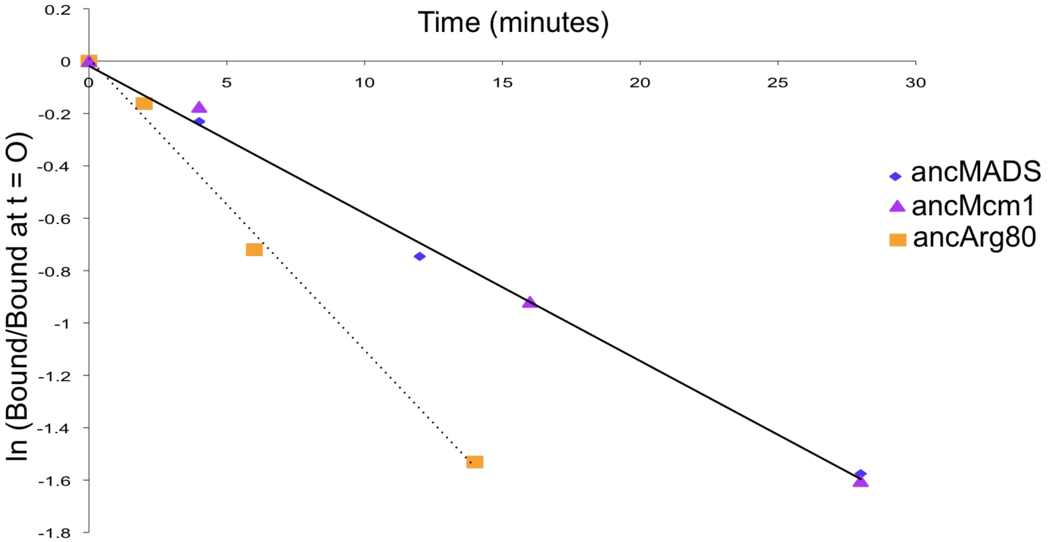


Figure 4— Reduced DNA-binding affinity of a post-duplication paralog

Quantifying the half-life of ancestral MADS-box proteins on the *S. cerevisiae* Arg3 *cis*-regulatory sequence. Unlabeled DNA was added at time point zero. The amount of protein bound to P^{32} labeled DNA was quantified using phosphoimaging.

Amar, N., F. Messenguy, et al. (2000). "ArgRII, a component of the ArgR-Mcm1 complex involved in the control of arginine metabolism in Saccharomyces cerevisiae, is the sensor of arginine." Mol Cell Biol **20**(6): 2087-2097.

Angel, P. and M. Karin (1991). "The role of Jun, Fos and the AP-1 complex in cell-proliferation and transformation." Biochim Biophys Acta **1072**(2-3): 129-157.

Baker, C. R., B. B. Tuch, et al. (2011). "Extensive DNA-binding specificity divergence of a conserved transcription regulator." Proc Natl Acad Sci U S A **108**(18): 7493-7498.

Bender, A. and G. F. Sprague, Jr. (1987). "MAT alpha 1 protein, a yeast transcription activator, binds synergistically with a second protein to a set of cell-type-specific genes." Cell **50**(5): 681-691.

Boonchird, C., F. Messenguy, et al. (1991). "Characterization of the yeast ARG5,6 gene: determination of the nucleotide sequence, analysis of the control region and of ARG5,6 transcript." Mol Gen Genet **226**(1-2): 154-166.

Byrne, K. P. and K. H. Wolfe (2007). "Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication." Genetics **175**(3): 1341-1350.

Finnigan, G. C., V. Hanson-Smith, et al. (2012). "Evolution of increased complexity in a molecular machine." Nature **481**(7381): 360-364.

Gordon, J. L., D. Armisén, et al. (2011). "Evolutionary erosion of yeast sex chromosomes by mating-type switching accidents." Proc Natl Acad Sci USA **108**(50): 20024-20029.

Guindon, S. and O. Gascuel (2003). "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood." Syst Biol **52**(5): 696-704.

Hanson-Smith, V., B. Kolaczkowski, et al. (2010). "Robustness of ancestral sequence reconstruction to phylogenetic uncertainty." Mol Biol Evol **27**(9): 1988-1999.

Hayes, T. E., P. Sengupta, et al. (1988). "The human c-fos serum response factor and the yeast factors GRM/PRTF have related DNA-binding specificities." Genes Dev **2**(12B): 1713-1722.

Innan, H. and F. Kondrashov (2010). "The evolution of gene duplications: classifying and distinguishing between models." Nat Rev Genet **11**(2): 97-108.

Jamai, A., E. Dubois, et al. (2002). "Swapping functional specificity of a MADS box protein: residues required for Arg80 regulation of arginine metabolism." Mol Cell Biol **22**(16): 5741-5752.

Lander, G. C., E. Estrin, et al. (2012). "Complete subunit architecture of the proteasome regulatory particle." Nature **482**(7384): 186-191.

Leid, M., P. Kastner, et al. (1992). "Purification, cloning, and RXR identity of the HeLa cell factor with which RAR or TR heterodimerizes to bind target sequences efficiently." Cell **68**(2): 377-395.

Lohse, M. B., R. E. Zordan, et al. (2010). "Distinct class of DNA-binding domains is exemplified by a master regulator of phenotypic switching in Candida albicans." Proc Natl Acad Sci U S A **107**(32): 14105-14110.

Löytynoja, A. and N. Goldman (2005). "An algorithm for progressive multiple alignment of sequences with insertions." Proc Natl Acad Sci USA **102**(30): 10557-10562.

Lynch, M. and J. S. Conery (2003). "The origins of genome complexity." Science **302**(5649): 1401-1404.

Lynch, M. and A. Force (2000). "The probability of duplicate gene preservation by subfunctionalization." Genetics **154**(1): 459-473.

Mead, J., A. R. Bruning, et al. (2002). "Interactions of the Mcm1 MADS box protein with cofactors that regulate mating in yeast." Mol Cell Biol **22**(13): 4607-4621.

Messenguy, F. and E. Dubois (1993). "Genetic evidence for a role for MCM1 in the regulation of arginine metabolism in Saccharomyces cerevisiae." Mol Cell Biol **13**(4): 2586-2592.

Mitrovich, Q. M., B. B. Tuch, et al. (2007). "Computational and experimental approaches double the number of known introns in the pathogenic yeast Candida albicans." Genome Res **17**(4): 492-502.

Qiu, H. F., E. Dubois, et al. (1990). "Functional analysis of ARGRI and ARGRIII regulatory proteins involved in the regulation of arginine metabolism in Saccharomyces cerevisiae." Mol Gen Genet **222**(2-3): 192-200.

Shore, P. and A. D. Sharrocks (1995). "The MADS-box family of transcription factors." Eur J Biochem **229**(1): 1-13.

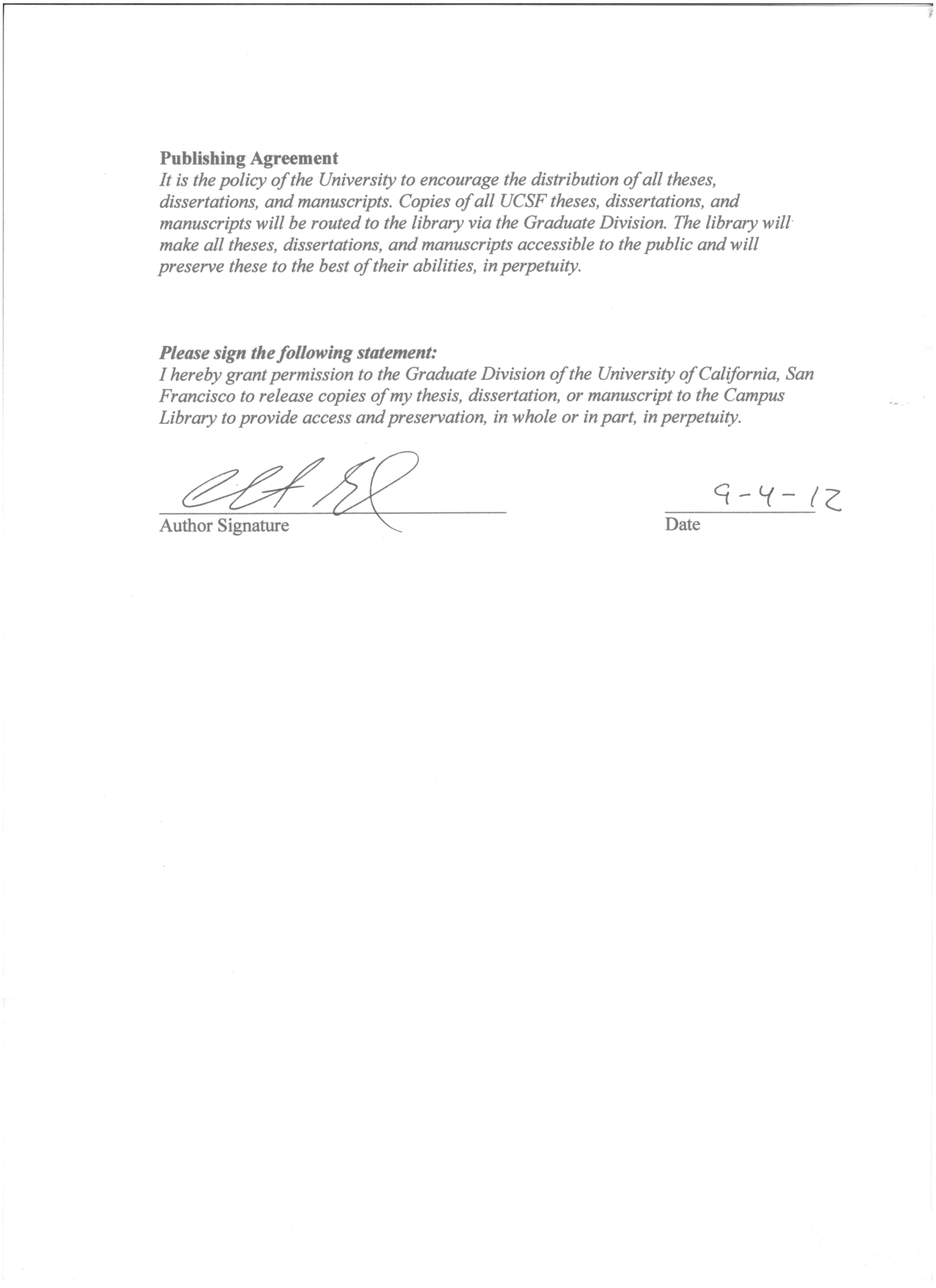
Tan, S. and T. J. Richmond (1998). "Crystal structure of the yeast MATalpha2/MCM1/DNA ternary complex." Nature **391**(6668): 660-666.

Taylor, J. W. and M. L. Berbee (2006). "Dating divergences in the Fungal Tree of Life: review and new analyses." Mycologia **98**(6): 838-849.

Thornton, J. W. (2004). "Resurrecting ancient genes: experimental analysis of extinct molecules." Nat Rev Genet **5**(5): 366-375.

Tuch, B. B., D. J. Galgoczy, et al. (2008). "The evolution of combinatorial gene regulation in fungi." PLoS Biol **6**(2): e38.

Zhang, L., B. S. Gaut, et al. (2001). "Gene duplication and evolution." Science **293**(5535): 1551.



Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature

9-4-12

Date